

LEWISBECK
BRYMAN
LIAO

A-F

1

SOCIAL SCIENCE RESEARCH METHODS



LEWISBECK
BRYMAN
LIAO

G-P

2

The SAGE Encyclopedia of
SOCIAL SCIENCE RESEARCH METHODS



LEWISBECK
BRYMAN
LIAO

Q-Z, Index

3

The SAGE Encyclopedia of
SOCIAL SCIENCE RESEARCH METHODS



The SAGE Encyclopedia of SOCIAL SCIENCE RESEARCH METHODS



VOLUME 3

MICHAEL S. LEWISBECK
ALAN E. BRYMAN
TIM FUTING LIAO
EDITORS



The SAGE Encyclopedia of
**SOCIAL SCIENCE
RESEARCH METHODS**

GENERAL EDITORS

Michael S. Lewis-Beck

Alan Bryman

Tim Futing Liao

EDITORIAL BOARD

Leona S. Aiken

Arizona State University, Tempe

R. Michael Alvarez

California Institute of Technology, Pasadena

Nathaniel Beck

University of California, San Diego, and New York University

Norman Blaikie

*Formerly at the University of Science,
Malaysia and the RMIT University, Australia*

Linda B. Bourque

University of California, Los Angeles

Marilynn B. Brewer

Ohio State University, Columbus

Derek Edwards

Loughborough University, United Kingdom

Floyd J. Fowler, Jr.

University of Massachusetts, Boston

Martyn Hammersley

The Open University, United Kingdom

Guillermina Jasso

New York University

David W. Kaplan

University of Delaware, Newark

Mary Maynard

University of York, United Kingdom

Janice M. Morse

University of Alberta, Canada

Martin Paldam

University of Aarhus, Denmark

Michael Quinn Patton

*The Union Institute & University,
Utilization-Focused Evaluation, Minnesota*

Ken Plummer

University of Essex, United Kingdom

Charles Tilly

Columbia University, New York

Jeroen K. Vermunt

Tilburg University, The Netherlands

Kazuo Yamaguchi

University of Chicago, Illinois

The SAGE Encyclopedia of
**SOCIAL SCIENCE
RESEARCH METHODS**

VOLUME 1

MICHAEL S. LEWIS-BECK

Department of Political Science, University of Iowa

ALAN BRYMAN

Department of Social Sciences, Loughborough University

TIM FUTING LIAO

Department of Sociology, University of Essex and University of Illinois, Urbana-Champaign

EDITORS

A SAGE Reference Publication



SAGE Publications

International Educational and Professional Publisher
Thousand Oaks ■ London ■ New Delhi

Copyright © 2004 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

For information:



SAGE Publications, Inc.
2455, Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
6, Bonhill Street
London EC2A 4PU
United Kingdom

SAGE Publications India Pvt. Ltd.
B-42, Panchsheel Enclave
Post Box 4109
New Delhi 110 017 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Lewis-Beck, Michael S.

The SAGE encyclopedia of social science research methods/Michael S.

Lewis-Beck, Alan Bryman, Tim Futing Liao.

p. cm.

Includes bibliographical references and index.

ISBN 0-7619-2363-2 (cloth)

1. Social sciences—Research—Encyclopedias. 2. Social sciences—Methodology—Encyclopedias.

I. Bryman, Alan. II. Liao, Tim Futing. III. Title.

H62.L456 2004

300!.! 72—dc22

2003015882

03 04 05 06 10 9 8 7 6 5 4 3 2 1

<i>Acquisitions Editor:</i>	Chris Rojek
<i>Publisher:</i>	Rolf A. Janke
<i>Developmental Editor:</i>	Eileen Haddox
<i>Editorial Assistant:</i>	Sara Tauber
<i>Production Editor:</i>	Melanie Birdsall
<i>Copy Editors:</i>	Gillian Dickens, Liann Lech, A. J. Sobczak
<i>Typesetter:</i>	C&M Digital (P) Ltd.
<i>Proofreader:</i>	Mattson Publishing Services, LLC
<i>Indexer:</i>	Sheila Bodell
<i>Cover Designer:</i>	Ravi Balasuriya

Contents

List of Entries, *vii*

Reader's Guide, *xix*

Contributors, *xxv*

About the Editors, *xli*

Preface, *xlili*

Introduction, *xlvi*

Entries

VOLUME I:

A–F

1–412

VOLUME II:

G–P

413–886

VOLUME III:

Q–Z

887–1208

Appendix: Bibliography, *1209*

Index, *1267*

List of Entries

Abduction
Access
Accounts
Action Research
Active Interview
Adjusted *R*-Squared. *See* Goodness-of-Fit Measures; Multiple Correlation; *R*-Squared
Adjustment
Aggregation
AIC. *See* Goodness-of-Fit Measures
Algorithm
Alpha, Significance Level of a Test
Alternative Hypothesis
AMOS (Analysis of Moment Structures)
Analysis of Covariance (ANCOVA)
Analysis of Variance (ANOVA)
Analytic Induction
Anonymity
Applied Qualitative Research
Applied Research
Archival Research
Archiving Qualitative Data
ARIMA
Arrow's Impossibility Theorem
Artifacts in Research Process
Artificial Intelligence
Artificial Neural Network. *See* Neural Network
Association
Association Model
Assumptions
Asymmetric Measures
Asymptotic Properties
ATLAS.ti
Attenuation
Attitude Measurement
Attribute
Attrition
Auditing
Authenticity Criteria
Autobiography
Autocorrelation. *See* Serial Correlation
Autoethnography
Autoregression. *See* Serial Correlation
Average
Bar Graph
Baseline
Basic Research
Bayes Factor
Bayes' Theorem, Bayes' Rule
Bayesian Inference
Bayesian Simulation. *See* Markov Chain Monte Carlo Methods
Behavior Coding
Behavioral Sciences
Behaviorism
Bell-Shaped Curve
Bernoulli
Best Linear Unbiased Estimator
Beta
Between-Group Differences
Between-Group Sum of Squares
Bias
BIC. *See* Goodness-of-Fit Measures
Bimodal
Binary
Binomial Distribution
Binomial Test
Biographic Narrative Interpretive Method (BNIM)
Biographical Method. *See* Life History Method
Biplots
Bipolar Scale
Biserial Correlation
Bivariate Analysis
Bivariate Regression. *See* Simple Correlation (Regression)
Block Design
BLUE. *See* Best Linear Unbiased Estimator
BMDP
Bonferroni Technique

- Bootstrapping
Bounded Rationality
Box-and-Whisker Plot. *See* Boxplot
Box-Jenkins Modeling
Boxplot

CAIC. *See* Goodness-of-Fit Measures
Canonical Correlation Analysis
CAPI. *See* Computer-Assisted Data Collection
Capture-Recapture
CAQDAS (Computer-Assisted Qualitative Data Analysis Software)
Carryover Effects
CART (Classification and Regression Trees)
Cartesian Coordinates
Case
Case Control Study
Case Study
CASI. *See* Computer-Assisted Data Collection
Catastrophe Theory
Categorical
Categorical Data Analysis
CATI. *See* Computer-Assisted Data Collection
Causal Mechanisms
Causal Modeling
Causality
CCA. *See* Canonical Correlation Analysis
Ceiling Effect
Cell
Censored Data
Censoring and Truncation
Census
Census Adjustment
Central Limit Theorem
Central Tendency. *See* Measures of Central Tendency
Centroid Method
Ceteris Paribus
CHAID
Change Scores
Chaos Theory
Chi-Square Distribution
Chi-Square Test
Chow Test
Classical Inference
Classification Trees. *See* CART
Clinical Research
Closed Question. *See* Closed-Ended Questions
Closed-Ended Questions
Cluster Analysis
Cluster Sampling
Cochran's Q Test
Code. *See* Coding
Codebook. *See* Coding Frame
Coding
Coding Frame
Coding Qualitative Data
Coefficient
Coefficient of Determination
Cognitive Anthropology
Cohort Analysis
Cointegration
Collinearity. *See* Multicollinearity
Column Association. *See* Association Model
Commonality Analysis
Communality
Comparative Method
Comparative Research
Comparison Group
Competing Risks
Complex Data Sets
Componential Analysis
Computer Simulation
Computer-Assisted Data Collection
Computer-Assisted Personal Interviewing
Concept. *See* Conceptualization, Operationalization, and Measurement
Conceptualization, Operationalization, and Measurement
Conditional Likelihood Ratio Test
Conditional Logit Model
Conditional Maximum Likelihood Estimation
Confidence Interval
Confidentiality
Confirmatory Factor Analysis
Confounding
Consequential Validity
Consistency. *See* Parameter Estimation
Constant
Constant Comparison
Construct
Construct Validity
Constructionism, Social
Constructivism
Content Analysis
Context Effects. *See* Order Effects
Contextual Effects
Contingency Coefficient. *See* Symmetric Measures
Contingency Table
Contingent Effects

-
- Continuous Variable
Contrast Coding
Control
Control Group
Convenience Sample
Conversation Analysis
Correlation
Correspondence Analysis
Counterbalancing
Counterfactual
Covariance
Covariance Structure
Covariate
Cover Story
Covert Research
Cramér's *V*. *See* Symmetric Measures
Creative Analytical Practice (CAP) Ethnography
Cressie-Read Statistic
Criterion Related Validity
Critical Discourse Analysis
Critical Ethnography
Critical Hermeneutics
Critical Incident Technique
Critical Pragmatism
Critical Race Theory
Critical Realism
Critical Theory
Critical Values
Cronbach's Alpha
Cross-Cultural Research
Cross-Lagged
Cross-Sectional Data
Cross-Sectional Design
Cross-Tabulation
Cumulative Frequency Polygon
Curvilinearity
- Danger in Research
Data
Data Archives
Data Management
Debriefing
Deception
Decile Range
Deconstructionism
Deduction
Degrees of Freedom
Deletion
Delphi Technique
Demand Characteristics
- Democratic Research and Evaluation
Demographic Methods
Dependent Interviewing
Dependent Observations
Dependent Variable
Dependent Variable (in Experimental Research)
Dependent Variable (in Nonexperimental Research)
Descriptive Statistics. *See* Univariate Analysis
Design Effects
Determinant
Determination, Coefficient of. *See* *R*-Squared
Determinism
Deterministic Model
Detrending
Deviant Case Analysis
Deviation
Diachronic
Diary
Dichotomous Variables
Difference of Means
Difference of Proportions
Difference Scores
Dimension
Disaggregation
Discourse Analysis
Discrete
Discriminant Analysis
Discriminant Validity
Disordinality
Dispersion
Dissimilarity
Distribution
Distribution-Free Statistics
Disturbance Term
Documents of Life
Documents, Types of
Double-Blind Procedure
Dual Scaling
Dummy Variable
Duration Analysis. *See* Event History Analysis
Durbin-Watson Statistic
Dyadic Analysis
Dynamic Modeling. *See* Simulation
- Ecological Fallacy
Ecological Validity
Econometrics
Effect Size
Effects Coding
Effects Coefficient

- Efficiency
Eigenvalues
Eigenvector. *See* Eigenvalues; Factor Analysis
Elaboration (Lazarsfeld's Method of)
Elasticity
Emic/Etic Distinction
Emotions Research
Empiricism
Empty Cell
Encoding/Decoding Model
Endogenous Variable
Epistemology
EQS
Error
Error Correction Models
Essentialism
Estimation
Estimator
Eta
Ethical Codes
Ethical Principles
Ethnograph
Ethnographic Content Analysis
Ethnographic Realism
Ethnographic Tales
Ethnography
Ethnomethodology
Ethnostatistics
Ethogenics
Ethology
Evaluation Apprehension
Evaluation Research
Event Count Models
Event History Analysis
Event Sampling
Exogenous Variable
Expectancy Effect. *See* Experimenter Expectancy Effect
Expected Frequency
Expected Value
Experiment
Experimental Design
Experimenter Expectancy Effect
Expert Systems
Explained Variance. *See* R-Squared
Explanation
Explanatory Variable
Exploratory Data Analysis
Exploratory Factor Analysis. *See* Factor Analysis
External Validity
Extrapolation
Extreme Case
F Distribution
F Ratio
Face Validity
Factor Analysis
Factorial Design
Factorial Survey Method (Rossi's Method)
Fallacy of Composition
Fallacy of Objectivism
Fallibilism
Falsificationism
Feminist Epistemology
Feminist Ethnography
Feminist Research
Fiction and Research
Field Experimentation
Field Relations
Field Research
Fieldnotes
File Drawer Problem
Film and Video in Research
First-Order. *See* Order
Fishing Expedition
Fixed Effects Model
Floor Effect
Focus Group
Focused Comparisons. *See* Structured, Focused Comparison
Focused Interviews
Forced-Choice Testing
Forecasting
Foreshadowing
Foucauldian Discourse Analysis
Fraud in Research
Free Association Interviewing
Frequency Distribution
Frequency Polygon
Friedman Test
Function
Fuzzy Set Theory

Game Theory
Gamma
Gauss-Markov Theorem. *See* BLUE; OLS
Geisteswissenschaften
Gender Issues
General Linear Models
Generalizability. *See* External Validity

-
- Generalizability Theory
 - Generalization/Generalizability in Qualitative Research
 - Generalized Additive Models
 - Generalized Estimating Equations
 - Generalized Least Squares
 - Generalized Linear Models
 - Geometric Distribution
 - Gibbs Sampling
 - Gini Coefficient
 - GLIM
 - Going Native
 - Goodness-of-Fit Measures
 - Grade of Membership Models
 - Granger Causality
 - Graphical Modeling
 - Grounded Theory
 - Group Interview
 - Grouped Data
 - Growth Curve Model
 - Guttman Scaling

 - Halo Effect
 - Hawthorne Effect
 - Hazard Rate
 - Hermeneutics
 - Heterogeneity
 - Heteroscedasticity. *See* Heteroskedasticity
 - Heteroskedasticity
 - Heuristic
 - Heuristic Inquiry
 - Hierarchical (Non)Linear Model
 - Hierarchy of Credibility
 - Higher-Order. *See* Order
 - Histogram
 - Historical Methods
 - Holding Constant. *See* Control
 - Homoscedasticity. *See* Homoskedasticity
 - Homoskedasticity
 - Humanism and Humanistic Research
 - Humanistic Coefficient
 - Hypothesis
 - Hypothesis Testing. *See* Significance Testing
 - Hypothetico-Deductive Method

 - Ideal Type
 - Idealism
 - Identification Problem
 - Idiographic/Nomothetic. *See* Nomothetic/Idiographic
 - Impact Assessment
 - Implicit Measures

 - Imputation
 - Independence
 - Independent Variable
 - Independent Variable (in Experimental Research)
 - Independent Variable (in Nonexperimental Research)
 - In-Depth Interview
 - Index
 - Indicator. *See* Index
 - Induction
 - Inequality Measurement
 - Inequality Process
 - Inference. *See* Statistical Inference
 - Inferential Statistics
 - Influential Cases
 - Influential Statistics
 - Informant Interviewing
 - Informed Consent
 - Instrumental Variable
 - Interaction. *See* Interaction Effect;
Statistical Interaction
 - Interaction Effect
 - Intercept
 - Internal Reliability
 - Internal Validity
 - Internet Surveys
 - Interpolation
 - Interpretative Repertoire
 - Interpretive Biography
 - Interpretive Interactionism
 - Interpretivism
 - Interquartile Range
 - Interrater Agreement
 - Interrater Reliability
 - Interrupted Time-Series Design
 - Interval
 - Intervening Variable. *See* Mediating
Variable
 - Intervention Analysis
 - Interview Guide
 - Interview Schedule
 - Interviewer Effects
 - Interviewer Training
 - Interviewing
 - Interviewing in Qualitative Research
 - Intraclass Correlation
 - Intracoder Reliability
 - Investigator Effects
 - In Vivo Coding
 - Isomorph
 - Item Response Theory

- Jackknife Method
- Key Informant
- Kish Grid
- Kolmogorov-Smirnov Test
- Kruskal-Wallis H Test
- Kurtosis
- Laboratory Experiment
- Lag Structure
- Lambda
- Latent Budget Analysis
- Latent Class Analysis
- Latent Markov Model
- Latent Profile Model
- Latent Trait Models
- Latent Variable
- Latin Square
- Law of Averages
- Law of Large Numbers
- Laws in Social Science
- Least Squares
- Least Squares Principle. *See* Regression
- Leaving the Field
- Leslie's Matrix
- Level of Analysis
- Level of Measurement
- Level of Significance
- Levene's Test
- Life Course Research
- Life History Interview. *See* Life Story Interview
- Life History Method
- Life Story Interview
- Life Table
- Likelihood Ratio Test
- Likert Scale
- LIMDEP
- Linear Dependency
- Linear Regression
- Linear Transformation
- Link Function
- LISREL
- Listwise Deletion
- Literature Review
- Lived Experience
- Local Independence
- Local Regression
- Loess. *See* Local Regression
- Logarithm
- Logic Model
- Logical Empiricism. *See* Logical Positivism
- Logical Positivism
- Logistic Regression
- Logit
- Logit Model
- Log-Linear Model
- Longitudinal Research
- Lowess. *See* Local Regression
- Macro
- Mail Questionnaire
- Main Effect
- Management Research
- MANCOVA. *See* Multivariate Analysis of Covariance
- Mann-Whitney U Test
- MANOVA. *See* Multivariate Analysis of Variance
- Marginal Effects
- Marginal Homogeneity
- Marginal Model
- Marginals
- Markov Chain
- Markov Chain Monte Carlo Methods
- Matching
- Matrix
- Matrix Algebra
- Maturation Effect
- Maximum Likelihood Estimation
- MCA. *See* Multiple Classification Analysis
- McNemar Change Test. *See* McNemar's Chi-Square Test
- McNemar's Chi-Square Test
- Mean
- Mean Square Error
- Mean Squares
- Measure. *See* Conceptualization, Operationalization, and Measurement
- Measure of Association
- Measures of Central Tendency
- Median
- Median Test
- Mediating Variable
- Member Validation and Check
- Membership Roles
- Memos, Memoing
- Meta-Analysis
- Meta-Ethnography
- Metaphors
- Method Variance
- Methodological Holism
- Methodological Individualism

-
- Metric Variable
 - Micro
 - Microsimulation
 - Middle-Range Theory
 - Milgram Experiments
 - Missing Data
 - Misspecification
 - Mixed Designs
 - Mixed-Effects Model
 - Mixed-Method Research. *See* Multimethod Research
 - Mixture Model (Finite Mixture Model)
 - MLE. *See* Maximum Likelihood Estimation
 - Mobility Table
 - Mode
 - Model
 - Model I ANOVA
 - Model II ANOVA. *See* Random-Effects Model
 - Model III ANOVA. *See* Mixed-Effects Model
 - Modeling
 - Moderating. *See* Moderating Variable
 - Moderating Variable
 - Moment
 - Moments in Qualitative Research
 - Monotonic
 - Monte Carlo Simulation
 - Mosaic Display
 - Mover-Stayer Models
 - Moving Average
 - Mplus
 - Multicollinearity
 - Multidimensional Scaling (MDS)
 - Multidimensionality
 - Multi-Item Measures
 - Multilevel Analysis
 - Multimethod-Multitrait Research. *See* Multimethod Research
 - Multimethod Research
 - Multinomial Distribution
 - Multinomial Logit
 - Multinomial Probit
 - Multiple Case Study
 - Multiple Classification Analysis (MCA)
 - Multiple Comparisons
 - Multiple Correlation
 - Multiple Correspondence Analysis.
See Correspondence Analysis
 - Multiple Regression Analysis
 - Multiple-Indicator Measures
 - Multiplicative
 - Multistage Sampling
 - Multi-Strategy Research
 - Multivariate
 - Multivariate Analysis
 - Multivariate Analysis of Variance and Covariance (MANOVA and MANCOVA)
 - N (n)
 - N6
 - Narrative Analysis
 - Narrative Interviewing
 - Native Research
 - Natural Experiment
 - Naturalism
 - Naturalistic Inquiry
 - Negative Binomial Distribution
 - Negative Case
 - Nested Design
 - Network Analysis
 - Neural Network
 - Nominal Variable
 - Nomothetic. *See* Nomothetic/Idiographic
 - Nomothetic/Idiographic
 - Nonadditive
 - Nonlinear Dynamics
 - Nonlinearity
 - Nonparametric Random-Effects Model
 - Nonparametric Regression. *See* Local Regression
 - Nonparametric Statistics
 - Nonparticipant Observation
 - Nonprobability Sampling
 - Nonrecursive
 - Nonresponse
 - Nonresponse Bias
 - Nonsampling Error
 - Normal Distribution
 - Normalization
 - NUD*IST
 - Nuisance Parameters
 - Null Hypothesis
 - Numeric Variable. *See* Metric Variable
 - NVivo
 - Objectivism
 - Objectivity
 - Oblique Rotation. *See* Rotations
 - Observation Schedule
 - Observation, Types of
 - Observational Research
 - Observed Frequencies
 - Observer Bias

- Odds
Odds Ratio
Official Statistics
Ogive. *See* Cumulative Frequency Polygon
OLS. *See* Ordinary Least Squares
Omega Squared
Omitted Variable
One-Sided Test. *See* One-Tailed Test
One-Tailed Test
One-Way ANOVA
Online Research Methods
Ontology, Ontological
Open Question. *See* Open-Ended Question
Open-Ended Question
Operational Definition. *See* Conceptualization, Operationalization, and Measurement
Operationism/Operationalism
Optimal Matching
Optimal Scaling
Oral History
Order
Order Effects
Ordinal Interaction
Ordinal Measure
Ordinary Least Squares (OLS)
Organizational Ethnography
Orthogonal Rotation. *See* Rotations
Other, the
Outlier
- p* Value
Pairwise Correlation. *See* Correlation
Pairwise Deletion
Panel
Panel Data Analysis
Paradigm
Parameter
Parameter Estimation
Part Correlation
Partial Correlation
Partial Least Squares Regression
Partial Regression Coefficient
Participant Observation
Participatory Action Research
Participatory Evaluation
Pascal Distribution
Path Analysis
Path Coefficient. *See* Path Analysis
Path Dependence
Path Diagram. *See* Path Analysis
- Pearson's Correlation Coefficient
Pearson's *r*. *See* Pearson's Correlation Coefficient
Percentage Frequency Distribution
Percentile
Period Effects
Periodicity
Permutation Test
Personal Documents. *See* Documents, Types of
Phenomenology
Phi Coefficient. *See* Symmetric Measures
Philosophy of Social Research
Philosophy of Social Science
Photographs in Research
Pie Chart
Piecewise Regression. *See* Spline Regression
Pilot Study
Placebo
Planned Comparisons. *See* Multiple Comparisons
Poetics
Point Estimate
Poisson Distribution
Poisson Regression
Policy-Oriented Research
Politics of Research
Polygon
Polynomial Equation
Polytomous Variable
Pooled Surveys
Population
Population Pyramid
Positivism
Post Hoc Comparison. *See* Multiple Comparisons
Postempiricism
Posterior Distribution
Postmodern Ethnography
Postmodernism
Poststructuralism
Power of a Test
Power Transformations. *See* Polynomial Equation
Pragmatism
Precoding
Predetermined Variable
Prediction
Prediction Equation
Predictor Variable
Pretest
Pretest Sensitization
Priming
Principal Components Analysis
Prior Distribution

-
- Prior Probability
 Prisoner's Dilemma
 Privacy. *See* Ethical Codes; Ethical Principles
 Privacy and Confidentiality
 Probabilistic
 Probabilistic Guttman Scaling
 Probability
 Probability Density Function
 Probability Sampling. *See* Random Sampling
 Probability Value
 Probing
 Probit Analysis
 Projective Techniques
 Proof Procedure
 Propensity Scores
 Proportional Reduction of Error (PRE)
 Protocol
 Proxy Reporting
 Proxy Variable
 Pseudo-*R*-Squared
 Psychoanalytic Methods
 Psychometrics
 Psychophysiological Measures
 Public Opinion Research
 Purposive Sampling
 Pygmalion Effect
- Q Methodology
 Q Sort. *See* Q Methodology
 Quadratic Equation
 Qualitative Content Analysis
 Qualitative Data Management
 Qualitative Evaluation
 Qualitative Meta-Analysis
 Qualitative Research
 Qualitative Variable
 Quantile
 Quantitative and Qualitative Research, Debate About
 Quantitative Research
 Quantitative Variable
 Quartile
 Quasi-Experiment
 Questionnaire
 Queuing Theory
 Quota Sample
- R*
r
 Random Assignment
 Random Digit Dialing
 Random Error
 Random Factor
 Random Number Generator
 Random Number Table
 Random Sampling
 Random Variable
 Random Variation
 Random Walk
 Random-Coefficient Model. *See* Mixed-Effects Model;
 Multilevel Analysis
 Random-Effects Model
 Randomized-Blocks Design
 Randomized Control Trial
 Randomized Response
 Randomness
 Range
 Rank Order
 Rapport
 Rasch Model
 Rate Standardization
 Ratio Scale
 Rationalism
 Raw Data
 RC(M) Model. *See* Association Model
 Reactive Measures. *See* Reactivity
 Reactivity
 Realism
 Reciprocal Relationship
 Recode
 Record-Check Studies
 Recursive
 Reductionism
 Reflexivity
 Regression
 Regression Coefficient
 Regression Diagnostics
 Regression Models for Ordinal Data
 Regression on . . .
 Regression Plane
 Regression Sum of Squares
 Regression Toward the Mean
 Regressor
 Relationship
 Relative Distribution Method
 Relative Frequency
 Relative Frequency Distribution
 Relative Variation (Measure of)
 Relativism
 Reliability
 Reliability and Validity in Qualitative Research

- Reliability Coefficient
Repeated Measures
Repertory Grid Technique
Replication
Replication/Replicability in Qualitative Research
Representation, Crisis of
Representative Sample
Research Design
Research Hypothesis
Research Management
Research Question
Residual
Residual Sum of Squares
Respondent
Respondent Validation. *See* Member Validation and Check
Response Bias
Response Effects
Response Set
Retroduction
Rhetoric
Rho
Robust
Robust Standard Errors
Role Playing
Root Mean Square
Rotated Factor. *See* Factor Analysis
Rotations
Rounding Error
Row Association. *See* Association Model
Row-Column Association. *See* Association Model
R-Squared

Sample
Sampling
Sampling Bias. *See* Sampling Error
Sampling Distribution
Sampling Error
Sampling Fraction
Sampling Frame
Sampling in Qualitative Research
Sampling Variability. *See* Sampling Error
Sampling With (or Without) Replacement
SAS
Saturated Model
Scale
Scaling
Scatterplot
Scheffé's Test
Scree Plot

Secondary Analysis of Qualitative Data
Secondary Analysis of Quantitative Data
Secondary Analysis of Survey Data
Secondary Data
Second-Order. *See* Order
Seemingly Unrelated Regression
Selection Bias
Self-Administered Questionnaire
Self-Completion Questionnaire. *See* Self-Administered Questionnaire
Self-Report Measure
Semantic Differential Scale
Semilogarithmic. *See* Logarithms
Semiotics
Semipartial Correlation
Semistructured Interview
Sensitive Topics, Researching
Sensitizing Concept
Sentence Completion Test
Sequential Analysis
Sequential Sampling
Serial Correlation
Shapiro-Wilk Test
Show Card
Sign Test
Significance Level. *See* Alpha, Significance Level of a Test
Significance Testing
Simple Additive Index. *See* Multi-Item Measures
Simple Correlation (Regression)
Simple Observation
Simple Random Sampling
Simulation
Simultaneous Equations
Skewed
Skewness. *See* Kurtosis
Slope
Smoothing
Snowball Sampling
Social Desirability Bias
Social Relations Model
Sociogram
Sociometry
Soft Systems Analysis
Solomon Four-Group Design
Sorting
Sparse Table
Spatial Regression
Spearman Correlation Coefficient
Spearman-Brown Formula

- Specification
Spectral Analysis
Sphericity Assumption
Spline Regression
Split-Half Reliability
Split-Plot Design
S-Plus
Spread
SPSS
Spurious Relationship
Stability Coefficient
Stable Population Model
Standard Deviation
Standard Error
Standard Error of the Estimate
Standard Scores
Standardized Regression Coefficients
Standardized Test
Standardized Variable. *See* z-Score; Standard Scores
Standpoint Epistemology
Stata
Statistic
Statistical Comparison
Statistical Control
Statistical Inference
Statistical Interaction
Statistical Package
Statistical Power
Statistical Significance
Stem-and-Leaf Display
Step Function. *See* Intervention Analysis
Stepwise Regression
Stochastic
Stratified Sample
Strength of Association
Structural Coefficient
Structural Equation Modeling
Structural Linguistics
Structuralism
Structured Interview
Structured Observation
Structured, Focused Comparison
Substantive Significance
Sum of Squared Errors
Sum of Squares
Summated Rating Scale. *See* Likert Scale
Suppression Effect
Survey
Survival Analysis
Symbolic Interactionism
Symmetric Measures
Symmetry
Synchronic
Systematic Error
Systematic Observation. *See* Structured Observation
Systematic Review
Systematic Sampling
Tau. *See* Symmetric Measures
Taxonomy
Tchebechev's Inequality
t-Distribution
Team Research
Telephone Survey
Testimonio
Test-Retest Reliability
Tetrachoric Correlation. *See* Symmetric Measures
Text
Theoretical Sampling
Theoretical Saturation
Theory
Thick Description
Third-Order. *See* Order
Threshold Effect
Thurstone Scaling
Time Diary. *See* Time Measurement
Time Measurement
Time-Series Cross-Section (TSCS) Models
Time-Series Data (Analysis/Design)
Tobit Analysis
Tolerance
Total Survey Design
Transcription, Transcript
Transfer Function. *See* Box-Jenkins Modeling
Transformations
Transition Rate
Treatment
Tree Diagram
Trend Analysis
Triangulation
Trimming
True Score
Truncation. *See* Censoring and Truncation
Trustworthiness Criteria
t-Ratio. *See* *t*-Test
t-Statistic. *See* *t*-Distribution
t-Test
Twenty Statements Test
Two-Sided Test. *See* Two-Tailed Test
Two-Stage Least Squares

- Two-Tailed Test
 Two-Way ANOVA
 Type I Error
 Type II Error
- Unbalanced Designs
 Unbiased
 Uncertainty
 Uniform Association. *See* Association Model
 Unimodal Distribution
 Unit of Analysis
 Unit Root
 Univariate Analysis
 Universe. *See* Population
 Unobserved Heterogeneity
 Unobtrusive Measures
 Unobtrusive Methods
 Unstandardized
 Unstructured Interview
 Utilization-Focused Evaluation
- V.* *See* Symmetric Measures
 Validity
 Variable
 Variable Parameter Models
 Variance. *See* Dispersion
 Variance Components Models
 Variance Inflation Factors
 Variation
 Variation (Coefficient of)
 Variation Ratio
 Varimax Rotation. *See* Rotations
 Vector
- Vector Autoregression (VAR). *See* Granger
 Causality
 Venn Diagram
 Verbal Protocol Analysis
 Verification
Verstehen
 Vignette Technique
 Virtual Ethnography
 Visual Research
 Volunteer Subjects
- Wald-Wolfowitz Test
 Weighted Least Squares
 Weighting
 Welch Test
 White Noise
 Wilcoxon Test
 WinMAX
 Within-Samples Sum of Squares
 Within-Subject Design
 Writing
- X* Variable
 $\bar{X}$
- Y*-Intercept
Y Variable
 Yates' Correction
 Yule's *Q*
- z*-Score
z-Test
 Zero-Order. *See* Order

The Appendix, Bibliography, appears at the end of this volume as well as in Volume 3.

Reader's Guide

ANALYSIS OF VARIANCE

Analysis of Covariance (ANCOVA)
Analysis of Variance (ANOVA)
Main Effect
Model I ANOVA
Model II ANOVA
Model III ANOVA
One-Way ANOVA
Two-Way ANOVA

ASSOCIATION AND CORRELATION

Association
Association Model
Asymmetric Measures
Biserial Correlation
Canonical Correlation Analysis
Correlation
Correspondence Analysis
Intraclass Correlation
Multiple Correlation
Part Correlation
Partial Correlation
Pearson's Correlation
 Coefficient
Semipartial Correlation
Simple Correlation (Regression)
Spearman Correlation Coefficient
Strength of Association
Symmetric Measures

BASIC QUALITATIVE RESEARCH

Autobiography
Life History Method
Life Story Interview
Qualitative Content Analysis
Qualitative Data Management
Qualitative Research

Quantitative and Qualitative Research,
 Debate About
Secondary Analysis of Qualitative Data

BASIC STATISTICS

Alternative Hypothesis
Average
Bar Graph
Bell-Shaped Curve
Bimodal
Case
Causal Modeling
Cell
Covariance
Cumulative Frequency Polygon
Data
Dependent Variable
Dispersion
Exploratory Data Analysis
F Ratio
Frequency Distribution
Histogram
Hypothesis
Independent Variable
Median
Measures of Central Tendency
N(*n*)
Null Hypothesis
Pie Chart
Regression
Standard Deviation
Statistic
t-Test
 $\bar{X}$
Y Variable
z-Test

CAUSAL MODELING

Causality
Dependent Variable

Effects Coefficient
Endogenous Variable
Exogenous Variable
Independent Variable
Path Analysis
Structural Equation Modeling

DISCOURSE/CONVERSATION ANALYSIS

Accounts
Conversation Analysis
Critical Discourse Analysis
Deviant Case Analysis
Discourse Analysis
Foucauldian Discourse Analysis
Interpretative Repertoire
Proof Procedure

ECONOMETRICS

ARIMA
Cointegration
Durbin-Watson Statistic
Econometrics
Fixed Effects Model
Mixed-Effects Model
Panel
Panel Data Analysis
Random-Effects Model
Selection Bias
Serial Correlation (Regression)
Time-Series Cross-Section (TSCS) Models
Time-Series Data (Analysis/Design)
Tobit Analysis

EPISTEMOLOGY

Constructionism, Social
Epistemology
Idealism
Interpretivism
Laws in Social Science
Logical Positivism
Methodological Holism
Naturalism
Objectivism
Positivism

ETHNOGRAPHY

Autoethnography
Case Study
Creative Analytical Practice (CAP) Ethnography

Critical Ethnography
Ethnographic Content Analysis
Ethnographic Realism
Ethnographic Tales
Ethnography
Participant Observation

EVALUATION

Applied Qualitative Research
Applied Research
Evaluation Research
Experiment
Heuristic Inquiry
Impact Assessment
Qualitative Evaluation
Randomized Control Trial

EVENT HISTORY ANALYSIS

Censoring and Truncation
Event History Analysis
Hazard Rate
Survival Analysis
Transition Rate

EXPERIMENTAL DESIGN

Experiment
Experimenter Expectancy Effect
External Validity
Field Experimentation
Hawthorne Effect
Internal Validity
Laboratory Experiment
Milgram Experiments
Quasi-Experiment

FACTOR ANALYSIS AND RELATED TECHNIQUES

Cluster Analysis
Commonality Analysis
Confirmatory Factor Analysis
Correspondence Analysis
Eigenvalue
Exploratory Factor Analysis
Factor Analysis
Oblique Rotation
Principal Components Analysis
Rotated Factor

Rotations
Varimax Rotation

FEMINIST METHODOLOGY

Feminist Ethnography
Feminist Research
Gender Issues
Standpoint Epistemology

GENERALIZED LINEAR MODELS

General Linear Models
Generalized Linear Models
Link Function
Logistic Regression
Logit
Logit Model
Poisson Regression
Probit Analysis

HISTORICAL/COMPARATIVE

Comparative Method
Comparative Research
Documents, Types of
Emic/Etic Distinction
Historical Methods
Oral History

INTERVIEWING IN QUALITATIVE RESEARCH

Biographic Narrative Interpretive
Method (BNIM)
Dependent Interviewing
Informant Interviewing
Interviewing in Qualitative Research
Narrative Interviewing
Semistructured Interview
Unstructured Interview

LATENT VARIABLE MODEL

Confirmatory Factor Analysis
Item Response Theory
Factor Analysis
Latent Budget Analysis
Latent Class Analysis
Latent Markov Model
Latent Profile Model

Latent Trait Models
Latent Variable
Local Independence
Nonparametric Random-Effects Model
Structural Equation Modeling

LIFE HISTORY/BIOGRAPHY

Autobiography
Biographic Narrative Interpretive
Method (BNIM)
Interpretive Biography
Life History Method
Life Story Interview
Narrative Analysis
Psychoanalytic Methods

LOG-LINEAR MODELS (CATEGORICAL DEPENDENT VARIABLES)

Association Model
Categorical Data Analysis
Contingency Table
Expected Frequency
Goodness-of-Fit Measures
Log-Linear Model
Marginal Model
Marginals
Mobility Table
Odds Ratio
Saturated Model
Sparse Table

LONGITUDINAL ANALYSIS

Cohort Analysis
Longitudinal Research
Panel
Period Effects
Time-Series Data (Analysis/Design)

MATHEMATICS AND FORMAL MODELS

Algorithm
Assumptions
Basic Research
Catastrophe Theory
Chaos Theory
Distribution
Fuzzy Set Theory
Game Theory

MEASUREMENT LEVEL

Attribute
Binary
Categorical
Continuous Variable
Dichotomous Variables
Discrete
Interval
Level of Measurement
Metric Variable
Nominal Variable
Ordinal Measure

**MEASUREMENT TESTING
AND CLASSIFICATION**

Conceptualization, Operationalization,
and Measurement
Generalizability Theory
Item Response Theory
Likert Scale
Multiple-Indicator Measures
Summated Rating Scale

MULTILEVEL ANALYSIS

Contextual Effects
Dependent Observations
Fixed Effects Model
Mixed-Effects Model
Multilevel Analysis
Nonparametric Random-Effects
Model
Random-Coefficient Model
Random-Effects Model

MULTIPLE REGRESSION

Adjusted *R*-Squared
Best Linear Unbiased Estimator
Beta
Generalized Least Squares
Heteroskedasticity
Interaction Effect
Misspecification
Multicollinearity
Multiple Regression Analysis
Nonadditive
R-Squared
Regression
Regression Diagnostics

Specification
Standard Error of the Estimate

QUALITATIVE DATA ANALYSIS

Analytic Induction
CAQDAS
Constant Comparison
Grounded Theory
In Vivo Coding
Memos, Memoing
Negative Case
Qualitative Content Analysis

SAMPLING IN QUALITATIVE RESEARCH

Purposive Sampling
Sampling in Qualitative Research
Snowball Sampling
Theoretical Sampling

SAMPLING IN SURVEYS

Multistage Sampling
Quota Sampling
Random Sampling
Representative Sample
Sampling
Sampling Error
Stratified Sampling
Systematic Sampling

SCALING

Attitude Measurement
Bipolar Scale
Dimension
Dual Scaling
Guttman Scaling
Index
Likert Scale
Multidimensional Scaling (MDS)
Optimal Scaling
Scale
Scaling
Semantic Differential Scale
Thurstone Scaling

SIGNIFICANCE TESTING

Alpha, Significance Level of a Test
Confidence Interval

Level of Significance
One-Tailed Test
Power of a Test
Significance Level
Significance Testing
Statistical Power
Statistical Significance
Substantive Significance
Two-Tailed Test

SIMPLE REGRESSION

Coefficient of Determination
Constant
Intercept
Least Squares
Linear Regression
Ordinary Least Squares
Regression on . . .
Regression
Scatterplot
Slope
Y-Intercept

SURVEY DESIGN

Computer-Assisted Personal
Interviewing
Internet Surveys

Interviewing
Mail Questionnaire
Secondary Analysis
of Survey Data
Structured Interview
Survey
Telephone Survey

TIME SERIES

ARIMA
Box-Jenkins Modeling
Cointegration
Detrending
Durbin-Watson Statistic
Error Correction Models
Forecasting
Granger Causality
Interrupted Time-Series Design
Intervention Analysis
Lag Structure
Moving Average
Periodicity
Serial Correlation
Spectral Analysis
Time-Series Cross-Section (TSCS)
Models
Time-Series Data (Analysis/Design)
Trend Analysis

Contributors

Abdi, Hervé

*University of Texas,
Dallas*

Adair, John G.

*University of Manitoba,
Canada*

Aditya, Ram N.

*Florida International University,
Miami*

Adler, Peter

*University of Denver,
Colorado*

Adler, Patricia A.

*University of Colorado,
Boulder*

Aiken, Leona S.

*Arizona State University,
Tempe*

Albritton, Robert B.

University of Mississippi

Aldridge, Alan

*University of Nottingham,
United Kingdom*

Allison, Paul D.

*University of Pennsylvania,
Philadelphia*

Altheide, David L.

*Arizona State University,
Tempe*

Altman, Micah

*Harvard University,
Massachusetts*

Alvesson, Mats

*University of Lund,
Sweden*

Amoureux, Jacques L.

University of Iowa, Iowa City

Andersen, Robert

*McMaster University,
Canada*

Anderson, Carolyn J.

*University of Illinois,
Urbana-Champaign*

Angle, John

*US Department of Agriculture,
Washington, DC*

Angrosino, Michael V.

*University of South Florida,
Tampa*

Anselin, Luc

*University of Illinois,
Urbana-Champaign*

Aronow, Edward

*Montclair State University,
New Jersey*

Ashmore, Malcolm

*Loughborough University,
United Kingdom*

Atkinson, Robert

*University of Southern Maine,
Gorham*

Atkinson, Rowland

*University of Glasgow,
United Kingdom*

Bacher, Johann

*Universität Erlangen-Nürnberg,
Germany*

Baird, Chardie L.

*Florida State University,
Tallahassee*

Bakeman, Roger

*Georgia State University,
Atlanta*

Bakker, Ryan

*University of Florida,
Gainesville*

Baltagi, Badi H.

Texas A&M University

Barr, Robert

*University of Manchester,
United Kingdom*

Bartunek, Jean M.

*Boston College,
Massachusetts*

Batson, C. Daniel

*University of Kansas,
Lawrence*

Beck, Nathaniel

*University of California, San Diego;
New York University*

Becker, Saul

*Loughborough University,
United Kingdom*

Becker, Mark P.

*University of Minnesota,
Minneapolis*

Benton, Ted

*University of Essex,
United Kingdom*

Berk, Richard A.

*University of California,
Los Angeles*

Billig, Michael

*Loughborough University,
United Kingdom*

Bishop, George F.

*University of Cincinnati,
Ohio*

Black, Thomas R.

*University of Surrey,
United Kingdom*

Blaikie, Norman

*Formerly at the University of Science,
Malaysia and the RMIT University,
Australia*

Blascovich, Jim

*University of California,
Santa Barbara*

Boehmke, Frederick J.

*University of Iowa,
Iowa City*

Boklund-Lagopoulos, Karin

*Aristotle University of Thessaloniki,
Greece*

Bollen, Kenneth A.

*University of North Carolina,
Chapel Hill*

Bornat, Joanna

*The Open University,
United Kingdom*

Bornstein, Robert F.

*Gettysburg College,
Pennsylvania*

Bottorff, Joan L.

*University of British Columbia,
Canada*

Bourque, Linda B.

*University of California,
Los Angeles*

Box-Steffensmeier, Janet M.

*Ohio State University,
Columbus*

Brady, Ivan

*State University of New York,
Oswego*

Brennan, Robert L.

*University of Iowa,
Iowa City*

Brewer, Marilynn B.

*Ohio State University,
Columbus*

Brick, J. Michael

*Westat, Rockville,
Maryland*

Brookfield, Stephen D.

*University of St. Thomas,
Minneapolis-St. Paul*

Brown, Courtney

*Emory University,
Atlanta*

Brown, Steven R.

Kent State University

Bryman, Alan

*Loughborough University,
United Kingdom*

Buck, Nicholas H.

*University of Essex,
United Kingdom*

Burkhart, Ross E.

Boise State University, Idaho

Burr, Vivien

*University of Huddersfield,
United Kingdom*

Byrne, David

*University of Durham,
United Kingdom*

Cameron, Amna

*North Carolina State University,
Raleigh*

Carless, Sally A.

*Monash University,
Australia*

Carmines, Edward G.

*Indiana University,
Bloomington*

Chapin, Tim

*Florida State University,
Tallahassee*

Charlton, Tony

*University of Gloucestershire,
United Kingdom*

Charlton, John R. H.

*Office for National Statistics,
United Kingdom*

Charmaz, Kathy

*Sonoma State University,
California*

Chartrand, Tanya L.

*Duke University,
North Carolina*

Chell, Elizabeth

*University of Southampton,
United Kingdom*

Chen, Peter Y.

*Colorado State University,
Fort Collins*

Christ, Sharon L.

*University of North Carolina,
Chapel Hill*

Ciambrone, Desirée

*Rhode Island College,
Providence*

Civettini, Andrew J.

*University of Iowa,
Iowa City*

Clark, M. H.

The University of Memphis, Tennessee

Clarke, Harold D.

*University of Texas,
Dallas*

Clegg, Chris

*University of Sheffield,
United Kingdom*

Clinton, Joshua D.

*Princeton University,
New Jersey*

Coffey, Amanda

*Cardiff University,
United Kingdom*

Cole, Michael

*University of California,
San Diego*

Cooper, Harris M.

*Duke University,
North Carolina*

Copp, Martha A.

*East Tennessee State University,
Johnson City*

Corbin, Juliet M.

*The University of Alberta,
Canada*

Corter, James E.

*Columbia University,
New York*

Corti, Louise

*University of Essex,
United Kingdom*

Cortina, Jose M.

*George Mason University,
Virginia*

Couper, Mick P.

*University of Michigan,
Ann Arbor*

Coxon, Anthony P. M.

*University of Edinburgh,
United Kingdom*

Craib, Ian

*Late of University of Essex,
United Kingdom*

Cramer, Duncan

*Loughborough University,
United Kingdom*

Crano, William D.

*Claremont Graduate University,
California*

Croll, Paul

*The University of Reading,
United Kingdom*

Croon, Marcel A.

*Tilburg University,
The Netherlands*

Currivan, Douglas B.

*University of Massachusetts,
Boston*

Dale, Angela

*University of Manchester,
United Kingdom*

De Boef, Suzanna

*Pennsylvania State University,
University Park*

de Vaus, David A.

*La Trobe University,
Australia*

Denzin, Norman K.

*University of Illinois,
Urbana-Champaign*

Dixon-Woods, Mary

*University of Leicester,
United Kingdom*

Doreian, Patrick

*London School of Economics,
United Kingdom*

Dovidio, John F.

*Colgate University,
New York*

Dressler, William W.

*The University of Alabama,
Tuscaloosa*

Drew, Paul

*University of York,
United Kingdom*

Easterby-Smith, Mark

*Lancaster University,
United Kingdom*

Edwards, Derek

*Loughborough University,
United Kingdom*

Edwards, W. Sherman

*Westat, Rockville,
Maryland*

Eisenhardt, Kathleen M.

*Stanford University,
California*

Eliason, Scott R.

*University of Minnesota,
Minneapolis*

Ellis, Charles H.

*Ohio State University,
Columbus*

Essed, Philomena

*University of Amsterdam,
The Netherlands*

Fairclough, Norman

*Lancaster University,
United Kingdom*

Fetterman, David M.

*Stanford University,
California*

Filmer, Paul

*Goldsmiths' College,
United Kingdom*

Finkel, Steven E.

*University of Virginia,
Charlottesville*

Flint, John

*University of Glasgow,
United Kingdom*

Foley, Hugh J.

*Skidmore College,
New York*

Fontana, Andrea

*University of Nevada,
Las Vegas*

Fowler, Jr., Floyd J.

*University of Massachusetts,
Boston*

Fox, James Alan

*Northeastern University,
Massachusetts*

Fox, John

*McMaster University,
Canada*

Franzosi, Roberto

*University of Reading,
United Kingdom*

Frederick, Richard I.

*US Medical Center for Federal Prisoners,
Missouri*

Freedman, David A.

*University of California,
Berkeley*

Frey, James H.

*University of Nevada,
Las Vegas*

Friendly, Michael

*York University,
Canada*

Fuller, Duncan

*University of Northumbria at Newcastle,
United Kingdom*

Furbee, N. Louanna

*University of Missouri,
Columbia*

Gallagher, Patricia M.

*University of Massachusetts,
Boston*

Gallmeier, Charles P.

*Indiana University Northwest,
Gary*

Garson, G. David

*North Carolina State University,
Raleigh*

Gephart, Jr., Robert P.

*University of Alberta,
Canada*

Gergen, Kenneth J.

*Swarthmore College,
Pennsylvania*

Gerhardt, Uta

*Universität Heidelberg,
Germany*

Gershuny, Jonathan

*University of Essex,
United Kingdom*

Gibbons, Jean D.

University of Alabama

Gibbs, Graham R.

*University of Huddersfield,
United Kingdom*

Gibson, Nancy

*University of Alberta,
Canada*

Gilbert, Nigel

*University of Surrey,
United Kingdom*

Gill, Jeff

*University of Florida,
Gainesville*

Glasgow, Garrett

*University of California,
Santa Barbara*

Glenn, Norval D.

*University of Texas,
Austin*

Goldstone, Jack

*University of California,
Davis*

Green, Donald P.

*Yale University,
Connecticut*

Greenacre, Michael

*Universitat Pompeu Fabra,
Spain*

Greenland, Sander

*University of California,
Los Angeles*

Greenwood, Davydd J.

*Cornell University,
New York*

Griffith, James W.

*Northwestern University,
Illinois*

Gross, Alan L.

City University of New York

Gschwend, Thomas

*University of Mannheim,
Germany*

Guba, Egon G.

Indiana University

Gubrium, Jaber F.

*University of Missouri,
Columbia*

Gujarati, Damodar N.

*United States Military Academy,
West Point*

Guo, Guang

*University of North Carolina,
Chapel Hill*

Guy, Will

*University of Bristol,
United Kingdom*

Hagenaars, Jacques A.

*Tilburg University,
The Netherlands*

Hagle, Timothy M.

*University of Iowa,
Iowa City*

Hammersley, Martyn

*The Open University,
United Kingdom*

Han, Shin-Kap

*University of Illinois,
Urbana-Champaign*

Hancock, Gregory R.

*University of Maryland,
College Park*

Handcock, Mark S.

*University of Washington,
Seattle*

Hansen, Peter Reinhard

Brown University, Rhode Island

Hardy, Melissa A.

*Pennsylvania State
University, State College*

Hargens, Lowell L.

*University of Washington,
Seattle*

Harper, Douglas

*Duquesne University,
Pennsylvania*

Harré, Rom

*University of Oxford,
United Kingdom*

Headland, Thomas N.

*SIL International, Dallas,
Texas*

Henry, Gary T.

*Georgia State University,
Atlanta*

Herron, Michael C.

*Dartmouth College,
New Hampshire*

Hessling, Robert M.

*University of Wisconsin,
Milwaukee*

Hine, Christine M.

*University of Surrey
United Kingdom*

Hipp, John R.

*University of North Carolina,
Chapel Hill*

Hocutt, Max

*The University of Alabama,
Tuscaloosa*

Hollway, Wendy

*The Open University,
United Kingdom*

Holman, Darryl J.

*University of Washington,
Seattle*

Holman, David J.

*University of Sheffield,
United Kingdom*

Holstein, James A.

*Marquette University,
Wisconsin*

Homan, Roger

*University of Brighton,
United Kingdom*

Horrace, William C.

*Syracuse University,
New York*

House, Ernest R.

*University of Colorado,
Boulder*

Hoyle, Rick H.

*Duke University,
North Carolina*

Hughes, Rhidian

*King's College London,
United Kingdom*

Hundley, Vanora

*University of Aberdeen,
United Kingdom*

Iversen, Gudmund R.

*Swarthmore College,
Pennsylvania*

Jaccard, James J.

*State University of New York,
Albany*

Jackman, Simon

*Stanford University,
California*

Jacoby, William G.

*Michigan State University,
East Lansing*

Jasso, Guillermina

New York University

Jefferis, Valerie E.

*Ohio State University,
Columbus*

Jefferson, Tony

*Keele University,
United Kingdom*

Jenkins, Richard

*University of Sheffield,
United Kingdom*

Johnson, Jeffrey C.

*East Carolina University,
Greenville*

Johnson, Paul E.

*University of Kansas,
Lawrence*

Jones, Michael Owen

*University of California,
Los Angeles*

Kanuha, Valli Kalei

*University of Hawaii,
Honolulu*

Kaplan, David W.

*University of Delaware,
Newark*

Karuntzos, Georgia T.

*RTI International, Research Triangle Park,
North Carolina*

Katz, Jonathan N.

*California Institute of Technology,
Pasadena*

Kedar, Orit

*University of Michigan,
Ann Arbor*

Kennedy, Peter E.

*Simon Fraser University,
Canada*

Kenny, David A.

*University of Connecticut,
Storrs*

Kimball, David C.

*University of Missouri,
St. Louis*

King, Jean A.

University of Minnesota, Minneapolis

Kirk, Roger E.

*Baylor University,
Texas*

Knoke, David H.

*University of Minnesota,
Minneapolis*

Kolen, Michael J.

*University of Iowa,
Iowa City*

Korczynski, Marek

*Loughborough University,
United Kingdom*

Kovtun, Mikhail

*Duke University,
North Carolina*

Krauss, Autumn D.

*Colorado State University,
Fort Collins*

Kroonenberg, Pieter M.

*Leiden University,
The Netherlands*

Krueger, Richard A.

*University of Minnesota,
St. Paul*

Kuhn, Randall

*University of Colorado,
Boulder*

Kuzel, Anton J.

*Virginia Commonwealth University,
Richmond*

Lagopoulos, Alexandros

*Aristotle University of Thessaloniki,
Greece*

Langeheine, Rolf

*Formerly at the University of Kiel,
Germany*

Laurie, Heather

*University of Essex,
United Kingdom*

Lauriola, Marco

*University of Rome,
"La Sapienza" Italy*

Lavrakas, Paul J.

*Nielsen Media Research,
New York*

Le Voi, Martin

*The Open University,
United Kingdom*

Leahey, Erin

*University of Arizona,
Tucson*

Ledolter, Johannes

*University of Iowa,
Iowa City*

Lee, Raymond M.

*Royal Holloway University
of London,
United Kingdom*

Lee (Zhen Li), Jane

*University of Florida,
Gainesville*

Levin, Irwin P.

*University of Iowa,
Iowa City*

Lewis-Beck, Michael S.

*University of Iowa,
Iowa City*

Liao, Tim Futing

*University of Essex,
United Kingdom;
University of Illinois,
Urbana-Champaign*

Lincoln, Yvonna S.

Texas A&M University

Little, Daniel

*University of Michigan,
Dearborn*

Little, Jani S.

*University of Colorado,
Boulder*

Liu, Cong*Illinois State University, Normal***Lockyer, Sharon***De Montfort University,
United Kingdom***Long, Karen J.***University of Michigan,
Ann Arbor***Lowe, Will***University of Bath,
United Kingdom***Lykken, David T.***University of Minnesota,
Minneapolis***Lynch, Michael***Cornell University,
New York***Lynn, Peter***University of Essex,
United Kingdom***Madsen, Douglas***University of Iowa,
Iowa City***Magidson, Jay***Statistical Innovations,
Belmont,
Massachusetts***Mahoney, Christine***Pennsylvania State University,
University Park***Mahoney, James***Brown University,
Rhode Island***Maines, David R.***Oakland University,
Michigan***Manton, Kenneth G.***Duke University,
North Carolina***Marsden, Peter V.***Harvard University,
Massachusetts***Marsh, Lawrence C.***University of Notre Dame,
Indiana***Mason, Jennifer***University of Leeds,
United Kingdom***Mathiowetz, Nancy***University of Wisconsin,
Milwaukee***Maxim, Paul S.***University of Western Ontario,
Canada***May, Tim***University of Salford,
United Kingdom***Maynard, Mary***University of York,
United Kingdom***McDowall, David***State University of New York,
Albany***McElhanon, Kenneth A.***Graduate Institute of Applied Linguistics,
Texas***McGraw, Kenneth O.***University of Mississippi***McGuigan, Jim***Loughborough University,
United Kingdom***McIver, John P.***University of Colorado,
Boulder***McKelvie, Stuart J.***Bishop's University,
Canada*

Meadows, Lynn M.
*University of Calgary,
Canada*

Menard, Scott
*University of Colorado,
Boulder*

Merrett, Mike C.
*University of Essex,
United Kingdom*

Mertens, Donna M.
*Gallaudet University,
Washington, DC*

Miller, Arthur G.
*Miami University,
Ohio*

Millward, Lynne
*University of Surrey,
United Kingdom*

Moffett, Kenneth W.
*University of Iowa,
Iowa City*

Mooney, Christopher Z.
*University of Illinois,
Springfield*

Moreno, Erika
*University of Iowa,
Iowa City*

Morris, Martina
University of Washington, Seattle

Morse, Janice M.
*University of Alberta,
Canada*

Moyé, Lemuel A.
*University of Texas,
Houston*

Mueller, Charles W.
*University of Iowa,
Iowa City*

Mueller, Daniel J.
*Indiana University,
Bloomington*

Mueller, Ralph O.
*The George Washington University,
Washington, DC*

Mukhopadhyay, Kausiki
*University of Denver,
Colorado*

Mulaik, Stanley A.
Georgia Institute of Technology, Atlanta

Nagler, Jonathan
New York University

Nash, Chris
*University of Technology,
Australia*

Nathan, Laura
*Mills College,
California*

Nishisato, Shizuhiko
*University of Toronto,
Canada*

Noblit, George W.
*University of North Carolina,
Chapel Hill*

Norpoth, Helmut
*State University of New York,
Stony Brook*

Noymer, Andrew
*University of California,
Berkeley*

Oakley, Ann
*University of London,
United Kingdom*

Olsen, Randall J.
Ohio State University, Columbus

Olsen, Wendy K.

*University of Manchester,
United Kingdom*

Ondercin, Heather L.

*Pennsylvania State University,
University Park*

Outhwaite, R. William

*University of Sussex,
United Kingdom*

Page, Stewart

*University of Windsor,
Canada*

Palmer, Harvey D.

University of Mississippi

Paluck, Elizabeth Levy

*Yale University,
Connecticut*

Pampel, Fred

*University of Colorado,
Boulder*

Parcero, Osiris J.

*Bristol University,
United Kingdom*

Park, Sung Ho

Texas A&M University

Parry, Ken W.

Griffith University, Australia

Pattison, Philippa

*University of Melbourne,
Australia*

Patton, Michael Quinn

*The Union Institute & University,
Utilization-Focused Evaluation, Minnesota*

Paul, Pallab

University of Denver, Colorado

Pavlichev, Alexei

*North Carolina State University,
Raleigh*

Paxton, Pamela

*Ohio State University,
Columbus*

Peng, Chao-Ying Joanne

*Indiana University,
Bloomington*

Phillips, Nelson

*University of Cambridge,
United Kingdom*

Phua, VoonChin

City University of New York

Pickering, Michael J.

*Loughborough University,
United Kingdom*

Pink, Sarah

*Loughborough University,
United Kingdom*

Platt, Lucinda

*University of Essex,
United Kingdom*

Platt, Jennifer

*University of Sussex,
United Kingdom*

Plummer, Ken

*University of Essex,
United Kingdom*

Poland, Blake D.

*University of Toronto,
Canada*

Pollins, Brian M.

*The Mershon Center, Columbus,
Ohio*

Post, Alison E.

*Harvard University,
Massachusetts*

Potter, Jonathan A.

*Loughborough University,
United Kingdom*

Prosch, Bernhard

*Universität Erlangen-Nürnberg,
Germany*

Pye, Paul E.

*University of Teesside,
United Kingdom*

Quinn, Kevin M.

*Harvard University,
Massachusetts*

Ragin, Charles C.

*University of Arizona,
Tucson*

Rässler, Susanne

*Universität Erlangen-Nürnberg,
Germany*

Redlawsk, David P.

*University of Iowa,
Iowa City*

Reis, Harry T.

*University of Rochester,
New York*

Richardson, Laurel

*Ohio State University,
Columbus*

Riessman, Catherine K.

*Boston University,
Massachusetts*

Ritzer, George

*University of Maryland,
College Park*

Robinson, Dawn T.

*University of Iowa,
Iowa City*

Rodgers, Joseph Lee

*University of Oklahoma,
Norman*

Rodríguez, Germán

*Princeton University,
New Jersey*

Roncek, Dennis W.

*University of Nebraska,
Omaha*

Rosenthal, Robert

*University of California,
Riverside*

Rosnow, Ralph L.

*Temple University,
Pennsylvania*

Rothstein, Hannah R.

*Baruch College,
New York*

Rubin, Donald B.

*Harvard University,
Massachusetts*

Rudas, Tamás

*Eotvos Lorand University,
Hungary*

Sandelowski, Margarete

*University of North Carolina,
Chapel Hill*

Santos, Filipe M.

*INSEAD,
France*

Sapsford, Roger

*University of Teesside,
United Kingdom*

Sasaki, Masamichi

*Hyogo Kyoiku University,
Japan*

Schafer, William D.

*University of Maryland,
College Park*

Schelin, Shannon Howle

*North Carolina State University,
Raleigh*

Schenker, Nathaniel

*National Center for Health Statistics,
Maryland*

Schmidt, Tara J.

*University of Wisconsin,
Milwaukee*

Schrodt, Philip A.

*University of Kansas,
Lawrence*

Scott, Steven L.

*University of Southern California,
Los Angeles*

Scott, John

*University of Essex,
United Kingdom*

Seale, Clive

*Brunel University,
United Kingdom*

Shadish, William R.

*University of California,
Merced*

Shaffir, William

*McMaster University,
Canada*

Shannon, Megan L.

*University of Iowa,
Iowa City*

Shipan, Charles R.

*University of Iowa,
Iowa City*

Shu, Xiaoling

*University of California,
Davis*

Sieber, Joan E.

*California State University,
Hayward*

Sijtsma, Klaas

*Tilburg University,
The Netherlands*

Silver, N. Clayton

*University of Nevada,
Las Vegas*

Slagter, Tracy Hoffmann

*University of Iowa,
Iowa City*

Smith, Peter W. F.

*University of Southampton,
United Kingdom*

Smith, John K.

*University of Northern Iowa,
Cedar Falls*

Smithson, Michael

*Australian National University,
Australia*

Snijders, Tom A. B.

*University of Groningen,
The Netherlands*

Spector, Paul E.

*University of South Florida,
Tampa*

Standing, Lionel G.

*Bishop's University,
Canada*

Stark, Philip B.

*University of California,
Berkeley*

Stevens, Gillian

*University of Illinois,
Urbana-Champaign*

Stillman, Todd

*University of Maryland,
College Park*

Stokes, S. Lynne

*Southern Methodist University,
Texas*

Stovel, Katherine

*University of Washington,
Seattle*

Strohmetz, David B.

*Monmouth University,
New Jersey*

Stuart, Elizabeth A.

*Harvard University,
Massachusetts*

Sutton, Alex

*University of Leicester,
United Kingdom*

Sweeney, Kevin J.

*Ohio State University,
Columbus*

Swicegood, C. Gray

*University of Illinois,
Urbana-Champaign*

Tacq, Jacques

*Catholic University of Brussels,
Belgium*

Tcherni, Maria

*Northeastern University,
Massachusetts*

Thomas, Jim

*Northern Illinois University,
DeKalb*

Thompson, Bruce

Texas A&M University

Thorne, Sally E.

*University of British Columbia,
Canada*

Tien, Charles

City University of New York

Ting, Kwok-fai

The Chinese University of Hong Kong

Toothaker, Larry E.

*University of Oklahoma,
Norman*

Tracy, Karen

*University of Colorado,
Boulder*

Traugott, Michael W.

*University of Michigan,
Ann Arbor*

Traxel, Nicole M.

*University of Wisconsin,
Milwaukee*

Tsuji, Ryuhei

*Meiji Gakuin University,
Japan*

Tuchman, Gaye

*University of Connecticut,
Storrs*

Tuma, Nancy Brandon

*Stanford University,
California*

Vallet, Louis-André

*LASMAS-Institut du Longitudinal (CNRS)
and Laboratoire de Sociologie
Quantitative (CREST),
France*

van de Pol, Frank J. R.

*Statistics Netherlands,
The Netherlands*

van der Ark, L. Andries

*Tilburg University,
The Netherlands*

van Manen, Max

*University of Alberta,
Canada*

van Teijlingen, Edwin R.

*University of Aberdeen,
United Kingdom*

Vann, Irvin B.

*North Carolina State University,
Raleigh*

Vermunt, Jeroen K.

*Tilburg University,
The Netherlands*

von Hippel, Paul T.

*Ohio State University,
Columbus*

Walker, Robert

*University of Nottingham,
United Kingdom*

Walsh, Susan

*University of Sheffield,
United Kingdom*

Wand, Jonathan

*Harvard University,
Massachusetts*

Wang, Cheng-Lung

*Florida State University,
Tallahassee*

Warren, Carol A. B.

*University of Kansas,
Lawrence*

Wasserman, Stanley

*University of Illinois,
Urbana-Champaign*

Weisberg, Herbert F.

*Ohio State University,
Columbus*

Weisfeld, Carol C.

*University of Detroit Mercy,
Michigan*

Wengraf, Tom

*Middlesex University,
United Kingdom*

Whitten, Guy

Texas A&M University

Wilcox, Rand R.

*University of Southern California,
Los Angeles*

Williams, Malcolm

*University of Plymouth,
United Kingdom*

Wood, B. Dan

Texas A&M University

Woods, James

*West Virginia University,
Morgantown*

Xie, Yu

*University of Michigan,
Ann Arbor*

Zimmerman, Marc

*University of Houston,
Texas*

Zorn, Christopher

*National Science Foundation,
Arlington, Virginia*

About the Editors

Michael S. Lewis-Beck is F. Wendell Miller Distinguished Professor of Political Science and Chair, at the University of Iowa. He has edited the SAGE green monograph series *Quantitative Applications in the Social Sciences* (QASS) since 1988. Also, he is Data Editor for the journal *French Politics*. He is past Editor of the *American Journal of Political Science*. Professor Lewis-Beck has authored or coauthored more than 110 books, articles, or chapters. His publications in methods include *Applied Regression: An Introduction* and *Data Analysis: An Introduction*. His publications in substantive areas of political science include *Economics and Elections: The Major Western Democracies*, *Forecasting Elections*, *How France Votes*, and *The French Voter: Before and After the 2002 Elections*. Besides the University of Iowa, he has taught quantitative methods courses at the Inter-university Consortium for Political and Social Research (ICPSR, University of Michigan); the European Consortium for Political Research (ECPR, University of Essex, England); the TARKI Summer School on Statistical Models (University of Budapest, Hungary); and the Ecole d'Eté de Lille (University of Lille, France). Also, Professor Lewis-Beck has held visiting appointments at the Institut d'Etudes Politiques (Sciences Po) in Paris; the Harvard University Department of Government; the University of Paris I (Sorbonne); the Catholic University in Lima, Peru; and the National Institute for Development Administration in Guatemala City. He speaks French and Spanish. His current research interests include election forecasting, and the modeling of coup behavior in Latin America.

Alan Bryman is Professor of Social Research in the Department of Social Sciences, Loughborough University, England. His main research interests lie in research methodology, leadership studies, organizational analysis, the process of Disneyization, human-animal relations, and theme parks. He is author or coauthor of many books, including *Quantity and*

Quality in Social Research (Routledge, 1988), *Charisma and Leadership in Organizations* (SAGE, 1992), *Disney and His Worlds* (Routledge, 1995), *Mediating Social Science* (SAGE, 1998), *Quantitative Data Analysis With SPSS Release 10 for Windows: A Guide for Social Scientists* (Routledge, 2001), *Social Research Methods* (Oxford University Press, 2001; rev. ed., 2004), *Business Research Methods* (Oxford University Press, 2003), and *Disneyization of Society* (SAGE, forthcoming). He is also coeditor of *The Handbook of Data Analysis* (SAGE, forthcoming). He is the editor of the *Understanding Social Research* series for Open University Press and is on the editorial board of *Leadership Quarterly*.

Tim Futing Liao is Professor of Sociology and Statistics at the University of Illinois, Urbana-Champaign, and currently (2001–2003) teaches research methods in the Department of Sociology at the University of Essex, UK. His research focuses on methodology, demography, and comparative and historical family studies. He has authored *Interpreting Probability Models* (SAGE, 1994), *Statistical Group Comparison* (Wiley, 2002), and research articles in both methods and disciplinary journals. Professor Liao served two terms (1992–1996 & 1996–2000) as Deputy Editor of *The Sociological Quarterly*, and he is on the Editorial Board of *Sociological Methods & Research* and the Steering Committee of the ESRC Oxford Spring School in Quantitative Methods for Social Research. His substantive interest led him to a year as visiting research fellow with the Cambridge Group for the History of Population and Social Structure, and his methodological expertise led him to teach summer schools at Peking University, Chinese Academy of Social Sciences, and the University of Essex. Having a catholic interest in methodology as well as in life, he has lived in or visited more than 20 nations, has done research on five societies, has held formal academic appointments in three educational systems, and speaks three and three-quarter languages.

Preface

These volumes comprise an encyclopedia of social science research methods, the first of its kind. Uniqueness explains, at least partly, why we undertook the project. It has never been done before. We also believe that such an encyclopedia is needed. What is an encyclopedia? In ancient Greek, the word signifies “all-encompassing education.” This reference set provides readers with an all-encompassing education in the ways of social science researchers. Why is this needed? No one researcher, expert or not, knows everything. He or she occasionally must look up some methodological point, for teaching as well as for research purposes. And it is not always obvious where to go, or what to look up when you get there. This encyclopedia brings together, in one place, authoritative essays on virtually all social science methods topics, both quantitative and qualitative.

A survey researcher may want to learn about longitudinal analysis in order to make sense of panel data. A student of household division of labor may want to learn the use of time diaries. A regression expert might bone up on multidimensional scaling, so as to be better able to aggregate certain measures. A deconstructionist might want to explore Foucauldian discourse analysis. An experimentalist may wish further understanding of the collinearity problem that nonexperimentalists routinely face. A feminist scholar may seek to know more about the use of films in research. A political scientist could have need of a refresher on experimental design. An anthropologist could raise a question about ethnographic realism. Perhaps a psychologist seeks to comprehend a particular measure of association. A philosopher might want to read more about the laws of social science. A sociologist may desire a review of the different approaches to evaluation research. Even in areas where one is highly expert, much can be gained. For example, in the examination of interaction effects, there are many entries from different perspectives, treating the little-used as well as the much-used techniques. These entries are written by acknowledged

leading scholars. One would have to be “more expert” than the experts not to benefit from reading them.

Besides practicing researchers and social statisticians, the encyclopedia has appeal for more general readers. Who might these people be? First, there are students, graduate and undergraduate, in the social science classes in the universities of the world. All the major branches—anthropology, communications, economics, education, geography, political science, psychology, public health, public policy, sociology, urban planning—have their students reading articles and books that require an appreciation of method for their full comprehension. The typical undergraduate needs to know how to read a contingency table, interpret a correlation coefficient, or critique a survey. There are entries on these and other such topics that are completely accessible to this student population. These sorts of entries require *no special knowledge of mathematics or statistics* to be understood, nor does the student need to have actually done research in order to grasp the essentials.

Of course, graduate students in these fields are obliged to go beyond simple reading comprehension, to actual application of these techniques. Many of the pieces make the “how-to” of the method clear, such as some of the entries on regression analysis or interviewing techniques. Then, there are more advanced entries that will challenge, but still inform, the typical graduate student. For example, the pieces on Bayesian analysis, or generalized linear modeling, will certainly allow them to sharpen their academic teeth.

In addition to researchers and students, this encyclopedia will be helpful to college-educated readers who want to understand more about social science methodology because of their work (e.g., they are journalists or city managers) or because of their pleasure (e.g., they follow social science reporting in the media). Many entries are written in ordinary English, with no special math or stat requirements, and so will be accessible to such readers. Moreover, many of the more difficult

entries are understandable in at least an introductory way, because of instructions to that effect given to our contributors. Having said that, it should be emphasized that none of the essays is “dumbed down.” Instead, they are intelligently written by recognized experts. There is nothing “cookbook” or condescending about them. In this domain of style, the only essential restriction placed on the authors was that of length, for we urged them to present their material as concisely as possible.

Our instructions to authors provided four different lengths for the entries—2,500 words (A-level), 1,000 words (B-level), 500 words (C-level), and 50 words (D-level). Of course, certain individual entries—whether A, B, C, or D level—do not conform exactly to these word counts, which are rather more guidelines than rigid standards. Length of entry was finally dictated by the importance, breadth, or, in some cases, complexity of the topic. In general, the longer the entry, the more central the topic. Examples of A-level entries are Analysis of Variance, Bayesian Inference, Case Study, and Regression. B-level topics are also important, but do not demand quite so much space to explain the concept clearly. B-level topics include Historical Method, Inferential Statistics, Misspecification, and Mixed Design. C-level entries usually treat a rather specific method or issue, such as Curvilinearity, Narrative Interview, Prisoner’s Dilemma, and Record-Check Studies. The A, B, and C entries all end with references to guide the reader further. D-level entries, on the other hand, are brief and mainly definitional. Examples of D-level entries are Gini Coefficient, Nominal, Secondary Data, and Yates’s Correction.

Altogether, there are about 1,000 entries, covering quantitative and qualitative methods, as well as the connections between them. The entries are cross-referenced to each other, as appropriate. For example, take this entry—Discrete (see Categorical, Nominal, Attribute). The number of entries, “1,000,” is not magic, nor was it a goal, but it is the number to which we kept returning in our search for a master list of entries. At the beginning of the project, we aimed to come up with an exhaustive list of all possible methods terms. We consulted written materials—subfield dictionaries, statistics encyclopedias, methods journals, publisher flyers—and individuals—members of our distinguished international board of scholars, editorial boards of methodology journals, colleagues at our home institutions and elsewhere, and, last but not least, our graduate students. As the process unfolded, some entries were culled, some consolidated, some added. We have included every methodological term

we could think of that someone reading social science research results in a book, a journal, or a newspaper might come across. In the spirit of the first encyclopedia, *l’Encyclopédie*, by Diderot and d’Alembert in 1751, we did try “to examine everything” (*Il faut tout examiner*) that concerned our enterprise. Although something is undoubtedly “not examined,” it remains for our readers to tell us what it is.

Once the entry list was basically set, we began contacting potential contributors. We were heartened by the extent of cooperation we received from the social science community. Senior and junior scholars alike gave us their best, and did so promptly. Remarkably, with a handful of exceptions, everyone met their deadlines. This is one measure, we believe, of the value these busy researchers place on this encyclopedia. As can be seen from the bylines, they compose a varied mix from the relevant disciplines, holding important university and research posts around the world.

Besides the contributors, the board, and the colleagues and students of our home institutions, there are others who are responsible for the success of the encyclopedia. Chris Rojek, senior editor at SAGE London, has to be credited with the initial idea of the project. Without that, it would not have happened. On the US side, C. Deborah Laughton, former senior editor at SAGE California, was a dedicated and tireless supporter of the encyclopedia idea from the very beginning. Without her enthusiasm and insight, the final product would have lacked much. Rolf Janke, vice president at SAGE, was a central source of organization strength and encouragement from the home office. Also at SAGE, Vincent Burns, technical editor, was vital in helping piece together the Web site, the database, and the massive manuscript data files. As well, we should mention Melanie Birdsall, production editor, who patiently shepherded us through the alphabet soup of page proofs. Finally, to save the best for last, we must thank Eileen Haddox, our developmental editor. Eileen made it work by doing the detail work. She sent out the contracts, reminded the authors, logged in the contributions, and sent text and corrections back and forth, in the end making the whole corpus ready for the printer. Always, she labored with efficiency and good cheer. In those realms, she set standards that we, the editors of the encyclopedia, tried to achieve. Sometimes we did.

Michael S. Lewis-Beck

Alan Bryman

Tim Futing Liao

Introduction

Social science methodology can be said to have entered our consciousness by the late 19th century when Emile Durkheim penned *The Rules of Sociological Method*, laying the foundation for the systematic understanding and analysis of social phenomena. From that time, it has seen continuous growth characterized by an exponential development over the last few decades of the 20th century. The development came in terms of both the width and the depth of our methodological know-how. The broad range of methods applicable in the wide circle of social science disciplines, and the sophistication and level in the advancement of some analytical approaches and techniques, would have been unthinkable merely a few score years ago. FOCUS GROUP and THICK DESCRIPTION, for example, have virtually become lingua franca among many social science researchers, regardless of their disciplinary orientation. The methods for dealing with MISSING DATA and NONRESPONSE, to take other examples, have advanced so much that specialized workshops on such issues are a perennial favorite among analysts of SURVEY DATA.

It is only natural, at the beginning of a new century, for us to take stock of the entire spectrum of our social science methodological knowledge, even though it is impossible and impractical to include every method that has ever been used in the social sciences. The cardinal aim of this encyclopedia is to provide our readers—be they students, academics, or applied researchers—with an introduction to a vast array of research methods by giving an account of their purposes, principles, developments, and applications. The approximately 1,000 entries, many of which are extensive treatments of the topics and contain recent developments, can be of great use to the novice or the experienced researcher alike.

To accomplish this goal, we offer two major types of entries: Some contain only a definition of no longer than a paragraph or two. These give the reader a quick explanation of a methodological term. True to

the encyclopedic form, many other entries are topical treatments or essays that discuss—at varying lengths, often with examples and sometimes with graphics—the nature, the history, the application, and the implication of using a certain method. Most of these entries also give suggested readings and references for the reader to pursue a topic further. These are part and parcel of the encyclopedia and are invaluable for those who would like to delve into the wonder world of research methods. To help provide a more complete explanation than is often achieved within the scope of a single article, we employ small capital letters, such as those appearing in the first paragraph of this introduction, that refer the reader to related terms that are explained elsewhere in the encyclopedia.

With such a variety of specialized essays to write, we are fortunate to have been able to count on the support of our board members and authors, who contributed many a coherent introduction to a method with definitiveness and thoroughness, often with great flair as well. Sometimes, topics are treated in such a novel way that they are not only pleasurable but also thought-provoking to read. For instance, entries such as the essay on ECONOMETRICS by Professor Damodar Gujarati are a pleasant surprise. Rather than merely introducing the topic with the types of methods and models that econometricians use and nothing else, Gujarati takes us on a journey from the ordinary to the extraordinary. He begins with three quotations that illustrate the broad scope of econometrics; here the simple, usual approach of using quotations accomplishes the seemingly undoable task of defining the terrain on which econometricians work and play. He then walks us twice through the research process, from economic theory to data and models to analysis, once in principle and the second time with an example. Such a process is what many of us preach every day but seldom think of when writing an essay for an encyclopedia. Gujarati uses the ordinary process of going about economic research to achieve an extraordinary,

profound impact—an impact that will leave a reader thinking about, instead of just the methods and models, the fundamental purpose of econometrics. Entries like this give us knowledge and food for thought.

The diversity of our entries is also great. To take one of many possible contrasts, some of our entries deal with very philosophical issues, such as POSTSTRUCTURALISM, that might appear to be out of step with a set of volumes concerned with methods of research, whereas others discuss advanced statistical techniques that might similarly be viewed as not part of social science research methodology. However, we have taken the view that both are necessary. On the one hand, all researchers need to be aware of the EPISTEMOLOGICAL issues that influence both the nature of RESEARCH QUESTIONS and the ASSUMPTIONS that underpin aspects of the research process; on the other hand, we all need to be aware of the full panoply of ways of analyzing

quantitative data, so that the most appropriate and robust techniques can be applied in different situations. It is only when we are knowledgeable about the choices available to us—whether epistemological, statistical, or whatever—that we can completely develop our craft as social researchers.

Examples of excellent treatment of a whole host of topics abound in the following pages and volumes. By assembling into one encyclopedia entries of varied origin that serve different research purposes, we, the editors, hope that readers will come to appreciate the rich heritage of our social science methodology and, more importantly, will be able to benefit from this immense source of methodological expertise in advancing their research.

Michael S. Lewis-Beck

Alan Bryman

Tim Futing Liao

A

ABDUCTION

Abduction is the logic used to construct descriptions and explanations that are grounded in the everyday activities of, as well as in the language and meanings used by, social actors. Abduction refers to the process of moving from the way social actors describe their way of life to technical, social scientific descriptions of that social life. It has two stages: (a) describing these activities and meanings and (b) deriving categories and concepts that can form the basis of an understanding or an explanation of the problem at hand. Abduction is associated with INTERPRETIVISM.

The logic of abduction is used to discover why people do what they do by uncovering largely tacit, mutual knowledge and the symbolic meanings, motives, and rules that provide the orientations for their actions. Mutual knowledge is background knowledge that is largely unarticulated, constantly used and modified by social actors as they interact with each other, and produced and reproduced by them in the course of their lives together. It is the everyday beliefs and practices—the mundane and taken for granted—that have to be grasped and articulated by the social researcher to provide an understanding of these actions.

The concept of abduction has had limited use in philosophy and social science. Two writers, the philosopher C. S. Peirce (1931, 1934) and the sociologist David Willer (1967), have used it but with different meanings to that presented here; the latter is derived from Blaikie (1993). Although not using the concept of abduction, the logic was devised by Alfred Schütz (1963) and used by Jack Douglas, John

Rex, and, particularly, Anthony Giddens (1976) (for a review of these writers, see Blaikie, 1993, pp. 163–168, 176–193).

What Schütz, Douglas, Rex, and Giddens have in common is their belief that social science accounts of the social world must be derived, at least initially, from the accounts that social actors can give of the aspect of their world of interest to the social scientist. They differ, however, in what is done with these accounts. Some argue that reporting everyday accounts is all that is possible or necessary to understand social life. Others are prepared to turn these accounts into social scientific descriptions of the way of life of a particular social group (community or society), but they insist on keeping these descriptions tied closely to social actors' language. Such descriptions lend themselves to two other possibilities. The first is to bring some existing THEORY or perspective to bear on them, thus providing a social scientific interpretation or critique of that way of life. The second is to generate some kind of explanation using as ingredients the IDEAL TYPES that are derived from everyday accounts. There are disagreements on both of these latter possibilities—in the first case, about whether criticism of another way of life is legitimate and, in the second case, about what else can or should be added to the everyday account to form a theory.

Interpretive social scientists use some version of abduction with a variety of research methods. However, they rarely articulate their logic. Appropriate methods for using abductive logic are still being developed.

—Norman Blaikie

REFERENCES

- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Giddens, A. (1976). *New rules of sociological method*. London: Hutchinson.
- Peirce, C. S. (1931). *Collected papers* (Vols. 1–2; C. Hartshorne & P. Weiss, Eds.). Cambridge, MA: Harvard University Press.
- Peirce, C. S. (1934). *Collected papers* (Vols. 5–6, C. Hartshorne & P. Weiss, Eds.). Cambridge, MA: Harvard University Press.
- Schütz, A. (1963). Concept and theory formation in the social sciences. In M. A. Natanson (Ed.), *Philosophy of the social sciences* (pp. 231–249). New York: Random House.
- Willer, D. (1967). *Scientific sociology: Theory and method*. Englewood Cliffs, NJ: Prentice Hall.

ACCESS

Access refers to a researcher seeking entry to an environment in which primary research data may be unearthed or generated. Therefore, for researchers interested in analyzing primary data, whether qualitative or quantitative in nature, access is an inevitable stage in the research process. The ability to secure access is a necessary part of the researcher's craft. Despite this, the process of securing access is rarely systematically examined, with researchers often adopting ad hoc, opportunistic, and intuitive approaches to the negotiation of access. When examining the range of tactics that have been used by researchers in their attempts to secure access, it is useful to begin by considering what researchers are seeking access into.

ACCESS TO WHAT, AND VIA WHOM?

Access is often considered in relation to what have been termed *closed and open/public settings* (Bell, 1969). Closed settings are formally constituted settings, such as firms, hospitals, and churches, and their archives. Open/public settings are informal in nature. Research into street corner life or local music scenes is an example of research in open/public settings. A key step in securing access into both closed and open settings involves negotiation with *gatekeepers*. Gatekeepers are the people who, metaphorically, have the ability to open or close the gate to the researcher seeking access into the setting. A preliminary step in seeking access lies in identifying the gatekeepers. In formal, closed settings, the gatekeeper will tend to be a

person in a senior position, such as a senior executive. Gatekeepers in informal, open settings are less easy to clearly identify but tend to have status and respect within the setting. Some researchers go directly to the gatekeeper in their attempts to secure access. Others approach the gatekeeper indirectly by first obtaining a “champion” or “sponsor” in the setting who can help persuade the gatekeeper to look favorably on the researcher's request for access.

OBTAINING ACCESS AS TRUST FORMATION AND NEGOTIATION

The ability of researchers to inspire sponsors, central gatekeepers, or dispersed individuals to trust them is central to the process of securing access. Obtaining access can be seen as a specific example of the wider social problem of trust creation. We can, therefore, usefully turn to the literature on trust creation to allow us a better and more systematic understanding of the process of securing access. Korczynski (2000) delineates four main bases of trust that may underlie an exchange. The most common tactics adopted by researchers in negotiating access can be mapped against these categories:

- *Trust based on personal relations*. It is common for researchers to use friends and contacts who know them as trustworthy as a way to obtain access. These friends and contacts may act as a conduit to sponsors, may be sponsors, or may actually be gatekeepers in the research settings.

- *Trust based on knowledge of the other party's internal norms*. Researchers may try to generate trust by giving signals to the gatekeeper or individual that they are “decent” people. For instance, seeking access to respondents by putting a notice in a newspaper or magazine associated with a particular moral, social, or political outlook implicitly invites the reader to associate the researcher with that newspaper's or magazine's norms and values.

- *Trust based on an incentive or governance structure surrounding the exchange*. Researchers can help generate trust by highlighting that it is in the interests of the researchers not to betray such trust. For instance, researchers may give gatekeepers the right to veto the publication of research findings (although such an approach leaves the researcher in a potentially precarious position).

- *Trust based on systems or institutions.* Researchers can seek to generate trust by invoking trust that gatekeepers or other individuals might have in institutions to which the researchers are affiliated. For instance, researchers may stress the quality of a university at which they are employed and highlight relevant research that has been produced by that university in the recent past.

The process of obtaining access should also be seen as a process of negotiation (Horn, 1996). Researchers typically want wide and deep access into a setting, whereas gatekeepers and other concerned individuals may be concerned about potential costs of granting such access. These potential conflicts of interest are typically informally aired in face-to-face meetings. In these implicit negotiations, researchers often offer the gatekeepers or individuals a benefit that would accrue from access being granted. In formal settings, this might take the form of a written report of the findings.

Trust generating and negotiating tend to be highly context-specific processes. This may go some way to explaining why research access tends to be discussed with an emphasis on context-specific tactics used by researchers. Few attempts have been made to construct a more abstract understanding of the process of securing research access.

ACCESS AS AN ONGOING PROCESS

For many qualitative researchers, the process of negotiating access is an ever-present part of the research process. For instance, researchers who wish to individually interview members of an informal group usually must negotiate access with each person in turn and cannot rely on cooperation flowing easily from an informal gatekeeper's implicit sanctioning of the research. Moreover, in formal settings, although respondents may feel the obligation to be interviewed by a researcher, they can still be reticent in their answers. In each case, the researcher needs to generate trust within the ongoing process of negotiating access.

ACCESS AND REACTIVITY

REACTIVITY refers to the presence of the researcher affecting the research environment. The way in which access is negotiated may affect reactivity significantly.

For instance, in organizational research, if trust is only generated with high-level gatekeepers and not with lower-level staff who are included in the study, the mode of securing access is likely to significantly bias the research results. As Burgess (1984) notes, "Access influences the reliability and validity of the data that the researcher subsequently obtains" (p. 45).

—Marek Korczynski

REFERENCES

- Bell, C. (1969). A note on participant observation. *Sociology*, 3, 417–418.
- Burgess, R. (1984). *In the field*. London: Allen & Unwin.
- Horn, R. (1996). Negotiating research access to organizations. *Psychologist*, 9(12), 551–554.
- Korczynski, M. (2000). The political economy of trust. *Journal of Management Studies*, 37(1), 1–22.

ACCOUNTS

An account is a piece of talk, a way of explaining the self or close others to another person. In an influential essay, sociologists Scott and Lyman (1968) highlighted how talk is the fundamental medium through which people manage relations with others: Offering accounts is how people remedy a problem with another or, more generally, make reasonable any action. Accounts include two main types: (a) accounts *for* action and (b) accounts *of* action.

An account *for* action is a statement offered to explain inappropriate conduct. The inappropriate behavior may be a small act, such as forgetting to make a promised telephone call, being late, or making an insensitive joke, or it may be a large transgression such as taking money without asking, having an affair, or lying about an important matter. Accounting is also done by one person for another, as when an employee seeks to excuse his boss's rude behavior by appealing to its unusualness. Accounts for questionable conduct are typically in the form of excuses or justifications. In giving an excuse, communicators accept that the act they did was wrong but deny that they had full responsibility. A student who tells her teacher, "I'm sorry this paper is late. My hard disk crashed yesterday," is using an excuse. In justifications, in contrast, a speaker accepts full responsibility for an act but denies that the act was wrong. An example would be a teen who

responded to his parent's reprimand about being late by saying, "I'm old enough to make my own decisions. You have no right to tell me what to do." Excuses and justifications are alternative ways of accounting, but they may also be paired. A person may say to his partner, "I'm sorry I hollered. I've been under so much pressure lately, but what do you expect when I'm in the middle of a conference call?" This person is simultaneously excusing and justifying, as well as offering an apology (acknowledging an act's inappropriateness and expressing sorrow).

The second and broader type of account is an account *of* action. An account of action (reason giving) is a pervasive feature of communicative life. Whenever people give reasons for what they are doing, they are accounting. Accounting is occurring when a woman tells a friend why she married Tim, a coworker mentions why he is no longer eating red meat, or a husband explains to his wife why they should visit Portugal rather than Greece. Accounts *of* action are often in the form of descriptively detailed stories. One nonobvious feature of accounts, both *for* and *of* action, is what they tell us about a group or culture. When people give reasons for their actions (or thoughts), they are treating the actions as choices. Of note is that people do not account for what they regard as natural or inevitable. Accounts, then, are a window into a culture. A close look at who offers them, about what they are accounting for, and under which circumstances makes visible a culture's beliefs about what counts as reasonable conduct.

—Karen Tracy

REFERENCES

- Buttny, R., & Morris, G. H. (2001). Accounting. In W. P. Robinson & H. Giles (Eds.), *The new handbook of language and social psychology* (pp. 285–301). Chichester, England: Wiley.
- Scott, M. B., & Lyman, S. (1968). Accounts. *American Sociological Review*, 33, 46–62.
- Tracy, K. (2002). *Everyday talk: Building and reflecting identities*. New York: Guilford.

ACTION RESEARCH

Action research is a strategy for addressing research issues in partnership with local people. Defining characteristics of action research are collaboration, mutual education, and action for change. This approach

increases the validity of research by recognizing contextual factors within the research environment that are often overlooked with more structured approaches. Action researchers are sensitive to culture, gender, economic status, ability, and other factors that may influence research partners, results, and research communities.

Ethics is a dominant theme in action research; researchers create an intentional ethical stance that has implications for process, philosophical framework, and evaluation procedures. Ethical principles are usually negotiated at the outset by all partners to identify common values that will guide the research process. These can include consultation throughout the research process, the valuing of various kinds of knowledge, mutual respect, and negotiation of the research questions, methods, dissemination, and follow-on action.

Action research can encompass the entire research project: identification and recruitment of the research team, definition of the research question, selection of data collection methods, analysis and interpretation of data, dissemination of results, evaluation of results and process, and design of follow-on programs, services, or policy changes. Although the degree of collaboration and community involvement varies from one research project to another, action research can be "active" in many ways depending on the needs or wishes of participants. All partners, including community members, play an active role in the research process; the research process itself is a training venue that builds research capacity for all partners; local knowledge and research skills are shared within the team; and the research outcome is a program or policy. Ideally, action research is transformative for all participants.

HISTORY

Action research has its roots in the work of Kurt Lewin (1935/1959) and the subsequent discourse in sociology, anthropology, education, and organizational theory; it is also based on the development theories of the 1970s, which assume the equality and complementarity of the knowledge, skills, and experience of all partners. As the human rights movement has challenged traditional scientific research models, new models for collaborative research have emerged. These models have various names: action research, PARTICIPATORY ACTION RESEARCH (PAR), collaborative research, and emancipatory or empowerment or

community-based research, with considerable overlap among them. All models share a commitment to effecting positive social change by expanding the traditional research paradigm to include the voices of those most affected by the research; thus, “research subjects” can become active participants in the research process. Most models include an element of capacity building or training that contributes to the empowerment of the community, the appropriateness of the dissemination process, and the sustainability of action research results.

CASE EXAMPLE: THE CANADIAN TUBERCULOSIS PROJECT

The Canadian Tuberculosis (TB) Project was a 3-year project that addressed the sociocultural factors influencing the prevention and treatment of TB among the most affected people: foreign-born and aboriginal populations. Initiated by a nurse at the TB clinic in the province of Alberta, the need for this research was echoed by leaders within the affected groups. Beginning with the establishment of a community advisory committee (CAC), a set of guiding principles was negotiated. This process contributed to team building and the establishment of trust among the various players, including community, government, and academic partners. The CAC then worked with the administrative staff to recruit two representatives from each of the four aboriginal and six foreign-born communities. These 20 community research associates were trained in action research principles, interview techniques, and qualitative data analysis. They conducted interviews within four groups in each of their communities: those with active TB, those on prophylaxis, those who refused prophylaxis, and those with a more distant family history of TB in their country of origin or on aboriginal reserves. The community research associates, the CAC, and the academic staff then analyzed the interview data using both manual and electronic techniques. The multimethod and multi-investigator analysis strategy maximized the validity of the data and the research findings.

It became very clear that the action research process was valuable to the collaborative team, as members learned to respect each other’s ways of knowing about the transmission and treatment of TB. Furthermore, the result of the study confirmed that although the health care system is very effective in treating active tuberculosis, there are few prevention programs in

high-risk communities. In addition, health professionals frequently miss TB in their screening because it is not common in the general population in Western countries. The results were reviewed by the CAC and the community research associates to ensure consistency with the original ethical principles. This group then recommended the dissemination plan and the follow-on actions, including an information video, a community education nurse position, and TB fact sheets in their various languages.

LIMITATIONS TO ACTION RESEARCH

The action research approach has been criticized for potential lack of rigor and investigator bias. However, all research can suffer from these weaknesses, and only vigilance ensures high quality of the process and outcomes, regardless of the approach or methodology. Other challenges associated with action research include having longer timelines associated with these projects, sustaining commitment and funding over long periods of time, maintaining consistency with multiple and/or inexperienced researchers, educating funders to support this approach, sustaining a strong theoretical framework, and reducing participant turnover.

EVALUATION

The complexity of action research necessitates a carefully documented and analyzed audit trail. Inclusion of the full range of stakeholders in both formative and process evaluations can enhance the validity of the data and the evaluation process itself (Reason & Bradbury, 2001). These evaluations are most effective when integrated as a routine activity throughout the project. The rigor of the research is also enhanced when the methodology and project goals are measured by their relevance to the research context (Gibson, Gibson, & Macaulay, 2001). It is an important challenge for evaluation to accommodate the dynamic nature of action research while measuring outcomes related to initial goals and evaluating the consistency of adherence to negotiated ethical standards.

OUTCOMES AND ACTION

The process of the action research experience, including mutual learning and capacity building through training, is as valuable as the more tangible research results. Hence, dissemination strategies

should include innovative mechanisms such as posters that can be used by community members and academics; use of community-accessible media such as radio, television, and newsletters; town meetings; and academic articles (which are reviewed by all partners prior to publication to prevent unauthorized release or unintentional misrepresentation of community or stigmatization). Because of the educational goal of action research, talks or posters may be presented in multiple venues or languages and can serve as training vehicles themselves, as well as dissemination strategies. The action research process can create effective alliances, building bridges across timeworn chasms between local communities and academics, government, and organizations.

—Nancy Gibson

REFERENCES

- Cooke, B., & Kothari, U. (Eds.). (2001). *Participation: The new tyranny?* London: Zed.
- Gibson, N., Gibson, G., & Macaulay, A. C. (2001). Community-based research: Negotiating agendas and evaluating outcomes. In J. Morse, J. Swanson, & A. J. Kuzel (Eds.), *The nature of qualitative evidence* (pp. 160–182). Thousand Oaks, CA: Sage.
- Lewin, K. (1959). *A dynamic theory of personality: Selected papers* (D. K. Adams & K. E. Zener, Trans.). New York: McGraw-Hill. (Original work published 1935.)
- Reason, P., & Bradbury, H. (Eds.). (2001). *Action Research: Participative inquiry and practice*. London: Sage.
- Stringer, E. T. (1996). *Action Research: A handbook for practitioners*. Thousand Oaks, CA: Sage.

ACTIVE INTERVIEW

The term *active interview* does not denote a particular type of interview as much as it suggests a specific orientation to the interview process. It signals that all interviews are unavoidably active, interactionally and conversationally. The dynamic, communicative contingencies of the interview—any interview—literally incite respondents' opinions. Every interview is an occasion for constructing, not merely discovering or conveying, information (Holstein & Gubrium, 1995).

INTERVIEWING is always conversational, varying from highly structured, standardized, survey interviews to free-flowing informal exchanges. The

narratives produced may be as truncated as FORCED-CHOICE survey responses or as elaborate as oral life histories, but they are all activated and constructed in situ, through interview participants' talk.

Although most researchers acknowledge the interactional character of the interview, the technical literature on interviewing tends to focus on keeping that interaction in check. Guides to interviewing, especially those oriented to standardized surveys, are primarily concerned with maximizing the flow of valid, reliable information while minimizing distortions of what respondents know. From this perspective, the standardized interview conversation is a pipeline for transmitting knowledge, whereas nonstandardized interaction represents a potential source of contamination that should be minimized.

In contrast, Charles Briggs (1986) argues that interviews fundamentally, not incidentally, shape the form and content of what is said. Aaron Cicourel (1974) goes further, maintaining that interviews virtually impose particular ways of understanding and communicating reality on subjects' responses. The point is that interviewers are deeply and unavoidably implicated in creating the meanings that ostensibly reside within respondents. Indeed, both parties to the interview are necessarily and ineluctably *active*. Meaning is not merely elicited by apt questioning, nor simply transported through respondent replies; it is communicatively assembled in the interview encounter. Respondents are not so much repositories of experiential information as they are constructors of knowledge in collaboration with interviewers (Holstein & Gubrium, 1995).

Conceiving of the interview as active leads to important questions about the very possibility of collecting and analyzing information in the manner the traditional approach presupposes (see Gubrium & Holstein, 1997). It also prompts researchers to attend more to the ways in which knowledge is assembled within the interview than is usually the case in traditional approaches. In other words, understanding *how* the meaning-making process unfolds in the interview is as critical as apprehending *what* is substantively asked and conveyed. The *hows* of interviewing, of course, refer to the interactional, narrative procedures of knowledge production, not merely to interview techniques. The *whats* pertain to the issues guiding the interview, the content of questions, and the

substantive information communicated by the respondent. A dual interest in the *hows* and *whats* of meaning construction goes hand in hand with an appreciation for the constitutive activeness of the interview process.

—James A. Holstein and Jaber F. Gubrium

REFERENCES

- Briggs, C. (1986). *Learning how to ask*. Cambridge, UK: Cambridge University Press.
- Cicourel, A. V. (1974). *Theory and method in a critique of Argentine fertility*. New York: John Wiley.
- Gubrium, J. F., & Holstein, J. A. (1997). *The new language of qualitative method*. New York: Oxford University Press.
- Holstein, J. A., & Gubrium, J. F. (1995). *The active interview*. Thousand Oaks, CA: Sage.

ADJUSTED R-SQUARED. See
GOODNESS-OF-FIT MEASURES; MULTIPLE
CORRELATION; R-SQUARED

ADJUSTMENT

Adjustment is the process by which the anticipated effect of a variable in which there is no primary interest is quantitatively removed from the relationship of primary interest. There are several well-accepted procedures by which this is accomplished.

The need for adjustment derives from the observation that relationships between variables that capture the complete attention of an investigator also have relationships with other variables in which there is no direct interest. The relationship between ethnicity and death from acquired immunodeficiency syndrome (AIDS) may be of the greatest interest to the researcher, but it must be acknowledged that death rates from this disease are also influenced by several other characteristics of these patients (e.g., education, acculturation, and access to health care). A fundamental question that confronts the researcher is whether it is truly ethnicity that is related to AIDS deaths or whether ethnicity plays no real role in the explanation of AIDS deaths at all. In the second circumstance, ethnicity serves merely as a surrogate for combinations of these other variables that, among themselves, truly

determine the cumulative mortality rates. Thus, it becomes important to examine the relationship between these other explanatory variables (or covariates) and that of ethnicity and death from AIDS.

Three analytic methods are used to adjust for the effect of a variable: (a) CONTROL, (b) stratification, and (c) REGRESSION analysis. In the control method, the variable whose effect is to be adjusted for (the adjustor) is not allowed to vary in an important fashion. This is easily accomplished for DICHOTOMOUS VARIABLES. As an example, if it is anticipated that the relationship between prison sentence duration and age at time of sentencing is affected by gender, then the effect is removed by studying the relationship in only men. The effect of gender has been neutralized by fixing it at one and only one value.

The process of stratification adds a layer of complication to implementing the control procedure in the process of variable adjustment. Specifically, the stratification procedure uses the technique of control repetitively, analyzing the relationship between the variables of primary interest on a background in which the adjustor is fixed at one level and then fixed at another level. This process works very well when there are a relatively small number of adjustor variable levels. For example, as in the previous example, the analysis of prison sentence duration and age can be evaluated in (a) only men and (b) only women. This evaluation permits an examination of the role that gender plays in modifying the relationship between sentence duration and age. Statistical procedures permit an evaluation of the relationship between the variables of interest within each of the strata, as well as a global assessment of the relationship between the two variables of interest across all strata. If the adjusting variable is not dichotomous but continuous, then it can be divided into different strata in which the strata definitions are based on a range of values for the adjustor.

The procedure of adjustment through regression analysis is somewhat more complex but, when correctly implemented, can handle complicated adjustment operations satisfactorily. Regression analysis itself is an important tool in examining the strength of association between variables that the investigator believes may be linked in a relationship. However, this regression model must be carefully engineered to be accurate. We begin with a straightforward model in regression analysis, the simple straight-line model. Assume we have collected n pairs of observations (x_i, y_i) , $i = 1$ to n . The investigator's goal is to

demonstrate that individuals with different values of x will have different values of y in a pattern that is best predicted by a straight-line relationship. Write the simple linear regression model for regressing y on x as follows:

$$E[y_i] = \beta_0 + \beta_1 x_i. \quad (1)$$

For example, y could be the prison sentence length of a prisoner convicted of a crime, and x could be the age of the prisoner at the time of his or her sentencing. However, the investigator believes that the socioeconomic status (SES) of the prisoner's family also plays a role in determining the prisoner's sentence duration. Holding SES constant in this setting would not be a useful approach because it is a continuous variable with wide variability. Consider instead the following model in which prison sentence duration is regressed on both age and the SES simultaneously:

$$\begin{aligned} E[\text{sentence duration}] \\ = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{SES}). \end{aligned} \quad (2)$$

The commonly used LEAST SQUARES estimates for the parameters β_0 , β_1 , and β_2 are readily obtained from many available software packages. The estimate of β_1 , the parameter of interest, is obtained by carrying out three regressions. First, the dependent variable sentence duration is regressed on the adjustor SES. What is left over from this regression is the part of the prison sentence duration variable that is unexplained by SES. This unexplained portion is the prison sentence duration RESIDUAL. Next, the variable age is regressed on the adjustor SES, and the age residual is computed. Finally, the prison sentence duration residual is regressed on the age residual. The result of this final regression produces the estimate of β_1 in equation (2).

These computations are carried out automatically by most regression software. The model produces an estimate of the relationship between sentence duration and age in which the influence of SES is "removed." Specifically, the regression procedure isolated and identified the relationship between sentence duration and the SES adjustor and then removed the influence of the adjustor, repeating this procedure for age. Finally, it regressed the residual of prison sentence duration (i.e., what was left of sentence duration after removing the SES influence) on the residual of age. Thus, in the case of regression analysis, adjustment has a specific meaning that is different from the definition used when one either controls or stratifies the adjustor. In the

case of regression analysis, adjustment means that we isolate, identify, and remove the influences of the adjustor variable. When the explanatory variables are themselves discrete, ANALYSIS OF VARIANCE (ANOVA) produces this result. The ANALYSIS OF COVARIANCE (ANCOVA) provides the same information when the explanatory variables are composed of a mixture of discrete and continuous variables.

This process can be implemented when there is more than one adjustor variable in the regression model. In that case, the procedure provides estimates of the strength of association between the variables of interest after making simultaneous adjustments for the influence of a battery of other variables.

—Lemuel A. Moyé

REFERENCES

- Deming, W. E. (1997). *Statistical adjustment of data*. New York: Dover.
- Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic research: Principles and quantitative methods*. New York: Van Nostrand Reinhold.
- Neter, J., Wasserman, W., & Kutner, M. (1990). *Applied linear statistical models* (3rd ed.). Homewood, IL: Irwin.

AGGREGATION

Aggregation is the process by which data are collected at a different level from the level of greatest interest. Data are aggregated because they are more available at the group level. For example, crime statistics are available from urban centers but, in general, not from every citizen of the city. Attempts are then made to identify exposure-response relationships among different aggregates. The ECOLOGICAL FALLACY is the false conclusion that a result about groups, or aggregates, of individuals applies to the individuals within these aggregates. Because each aggregate contains individuals who differ from each other, the conclusion about the aggregate does not apply to each of its member individuals.

—Lemuel A. Moyé

See also ECOLOGICAL FALLACY

AIC. See GOODNESS-OF-FIT MEASURES

ALGORITHM

Algorithm is a computer science term for a way of solving a problem and also refers to the instructions given to the computer to solve the problem. The study of algorithms is central to computer science and is of great practical importance to data analysis in the social sciences because algorithms are used in statistical programs.

An algorithm can be thought of as any step-by-step procedure for solving a task. Imagine five playing cards face down on a table and the task of sorting them. Picking them up one at a time with the right hand and placing them in the left hand in their proper place would be one way to solve this task. This is an algorithm, called *insertion sort* in computer science.

It is worth noting the subtle distinction between the concept of an algorithm and the concept of a method or technique. For example, a method would be LEAST SQUARES, matrix inversion would be a technique used therein, and *LU decomposition* and *Strassen's algorithm* would be alternative algorithms to accomplish matrix inversion. A single data analysis method may use more than one algorithm.

It is impossible to write statistical software without using algorithms, so the importance of algorithms to quantitative methodologists is ensured. However, user-friendly statistical software packages eliminate the need for end users to construct their own algorithms for most tasks. Nonetheless, at least a basic understanding of algorithms can be useful. For example, MAXIMUM LIKELIHOOD ESTIMATION methods can use an initial estimate as a starting point, and in some cases, failure to converge may be remediated by trivially altering the initial estimate. Without some familiarity of the underlying algorithm, a researcher may be stuck with a nonconverging function.

Another setting where some knowledge of algorithms is useful is presented in Figure 1, which shows two possible digraph depictions of the same network data. The left panel uses the *multidimensional scaling* algorithm, and the right uses *simulated annealing*. The data are identical, which may be verified by observing who is connected to whom, but the appearance of the graphs is different. Algorithms are important here because interpretation of the network data is affected by the appearance of the graph, which is affected in turn by the choice of algorithm. Although in many cases, different algorithms will produce the same result

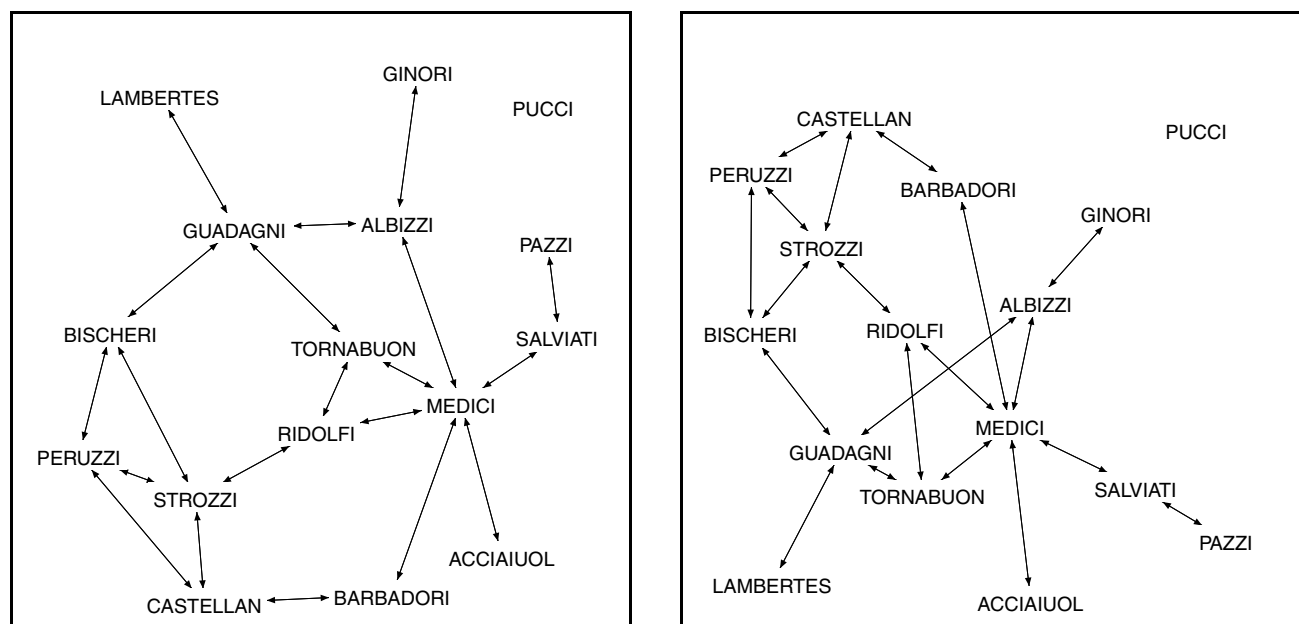


Figure 1 Two Possible Digraph Depictions of the Same Network Data

but differ in speed, in this case, different algorithms produce different results.

The term *algorithm* is sometimes used more broadly to mean any step-by-step procedure to solve a given task, whether or not a computer is involved. For instance, matching historical records from more than one archival source can be done by hand using an algorithm.

—Andrew Noymer

REFERENCES

- Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to algorithms* (2nd ed.). Cambridge, MA: MIT Press.
- Knuth, D. E. (1997). *Fundamental algorithms: The art of computer programming* (Vol. 1, 3rd ed.). Reading, MA: Addison-Wesley.

ALPHA, SIGNIFICANCE LEVEL OF A TEST

The development of SIGNIFICANCE TESTING is attributed to Ronald Fisher. His work in agrarian statistics in the 1920s led him to choose the significance threshold of .05, a level that has been accepted by research communities in many fields. The underlying motivations for the acceptance of hypothesis testing and the *P* VALUE have been reviewed (Goodman, 1999). However, it must be emphasized that no mathematical theory points to .05 as the optimum TYPE I ERROR level—only tradition.

The alpha error level is that level prospectively chosen during the design phase of the research; the *p* value is that error level computed based on the research data. Alpha levels are selected because they indicate the possibility that random sampling error has produced the result of the study, so they may be chosen at levels other than .05. The selection of an alpha level greater than .05 suggests that the investigators are willing to accept a greater level of Type I error to ensure the likelihood of identifying a positive effect in their research. Alternatively, the consequences of finding a positive effect in the research effort may be so important or cause such upheaval that the investigators set stringent requirements for the alpha error rate (e.g., $\alpha = .01$). Each of these decisions is appropriate if it (a) is made prospectively and (b) is adequately explained to the scientific community.

One adaptation of significance testing has been the use of ONE-TAILED TESTING. However, one-sided testing has been a subject of contention in medical research and has been the subject of controversy in the social sciences as well. An adaptation of TWO-TAILED TESTING, which leads to nontraditional alpha error levels, is the implementation of asymmetric regions of significance. When testing for the effect of an intervention in improving education in elementary school age children, it is possible that the intervention may have a paradoxical effect, reducing rather than increasing testing scores. In this situation, the investigator can divide the Type I error rate, so that 80% of the available Type I error level (.04 if the total alpha error level is .05) is in the “harm” end of the tail, and put the remaining 1% in the benefit tail of the distribution. The prospective declaration of such a procedure is critical and, when in place, demonstrates the investigator’s sensitivity to the possibility that the intervention may be unpredictably harmful. Guidelines for the development of such asymmetric testing have been produced (Moyé, 2000).

When investigators carry out several hypothesis tests within a clinical trial, the family-wise or overall Type I error level increases. A research effort that produces two statistical hypotheses resulting in *p* values at the .045 level may seem to suggest that the research sample has produced two results that are very likely to reflect findings in the larger population from which the sample was derived. However, an important issue is the likelihood that at least one of these conclusions is wrong. Assuming each of the two tests is independent, then the probability that at least one of these analyses produces a misleading result through sampling error alone is $1 - (.955)^2 = .088$. This can be controlled by making adjustments in the Type I error levels of the individual tests (e.g., BONFERRONI TECHNIQUES), which are useful in reducing the alpha error level for each statistical hypothesis test that was carried out. As an example, an investigator who is interested in examining the effect of a smoking cessation program on teenage smoking may place a Type I error rate of .035 on the effect of the intervention on self-reports of smoking and place the remaining .015 on the intervention’s effect on attitudes about smoking. This approach maintains the overall error rate at .05 while permitting a positive result for the effect of therapy on either of these two endpoint measurements of smoking.

—Lemuel A. Moyé

REFERENCES

- Goodman, S. N. (1999). Toward evidence-based medical statistics: 1. The p value fallacy. *Annals of Internal Medicine*, 130, 995–1004.
- Moyé, L. A. (2000). *Statistical reasoning in medicine*. New York: Springer-Verlag.

ALTERNATIVE HYPOTHESIS

The HYPOTHESIS that opposes the NULL HYPOTHESIS is commonly called the alternative. It is usually the research, or dominant, hypothesis. Suppose, for example, the research hypothesis is that sugar consumption is related to hyperactivity in children. That hypothesis, H_A , is an alternative to the null, H_0 , that sugar consumption is not related to hyperactivity.

—Michael S. Lewis-Beck

AMOS (ANALYSIS OF MOMENT STRUCTURES)

AMOS is a statistical software package for social sciences analyses applying STRUCTURAL EQUATION MODELS or COVARIANCE STRUCTURE MODELS. AMOS has good graphic interface. For more information, see the Web page of the software at www.spss.com/spssbi/amos/.

—Tim Futing Liao

ANALYSIS OF COVARIANCE (ANCOVA)

The methodology with this name grew out of a desire to combine ANALYSIS OF VARIANCE and REGRESSION analysis. It received considerable interest before the arrival of good computer packages for statistics, but the separate name for this methodology is now in decreasing use. More often, the analysis is simply referred to as multiple regression analysis, with both QUANTITATIVE VARIABLES and DUMMY VARIABLES as explanatory variables on a quantitative dependent variable.

DEFINITION

Analysis of covariance is a multivariate statistical method in which the dependent variable is a quantitative variable and the independent variables are a mixture of NOMINAL VARIABLES and quantitative variables.

EXPLANATION

Analysis of variance is used for studying the relationship between a quantitative dependent variable and one or more nominal independent variables. Regression analysis is used for studying the relationship between a quantitative dependent variable and one or more quantitative independent variables. This immediately raises the question of what to do when, in a common situation, the research hypothesis calls for some of the independent variables to be nominal and some to be quantitative variables.

With the arrival of statistical computer packages, it soon became clear that such a situation can easily be covered by introducing one or more dummy variables for the nominal variables and doing an ordinary multiple regression analysis. This can be done with or without the construction of INTERACTION variables, depending on the nature of the available data and the problem at hand. The use of a separate name for analysis of covariance has been decreasing because regression analysis can handle this analysis without any difficulties.

HISTORICAL ACCOUNT

A good deal of work on the development of analysis of covariance took place in the 1930s. As with so much of analysis of variance, Sir Ronald A. Fisher had a leading role in its development, which was highlighted in a special issue of the *Biometrics* journal, with an introductory paper by William G. Cochran (1957) and six other papers by leading statisticians of the time.

Analysis of covariance was a natural outgrowth of both analysis of variance and regression analysis. In analysis of variance, it may be reasonable to control for a quantitative variable and compare adjusted means instead of the raw, group means when comparing groups. Similarly, in regression analysis, it may be reasonable to control for a nominal variable. Both of these desires resulted in developments of analysis of covariance and shortcut methods for computations, using desk calculators.

APPLICATIONS

In its simplest form, analysis of covariance makes use of one quantitative variable and one nominal variable as explanatory variables. In regression notation, the model can be written as

$$Y = \alpha + \beta_1 X + \beta_2 D, \quad (1)$$

where Y and X are quantitative variables; D is a dummy variable with values 0 and 1, representing two groups; and α , β_1 , and β_2 are the regression parameters.

It is now possible to do analysis of variance comparing the means for the two groups while controlling for the X variable. Instead of comparing the means of Y for the two groups directly, we compare *adjusted* means of Y . By substituting 0 and 1 for the dummy variable, we get the two parallel lines with the following equations:

$$Y_1 = \alpha + \beta_1 x_1, \quad (2)$$

$$Y_2 = (\alpha + \beta_2) + \beta_1 x_2. \quad (3)$$

For a common value of the X variable—say, the overall mean—the difference between the predicted Y values becomes the value of the coefficient β_2 . Thus, when we control for the X variable by holding it constant at a fixed value, the adjusted difference in Y means becomes the vertical difference between the two lines above.

It is also possible to compare the two groups and control for the nominal group variable D . We could do that by dividing the data up into two groups and run one regression of Y on X for the observations where $D = 0$ and another for $D = 1$. This can be done when the two groups have identical slopes, and statisticians have developed elaborate computing formulas for this case using only desk calculators. But it would be more efficient to run one multivariate regression analysis with two independent variables instead of two separate, simple regressions.

The method generalizes to the use of more than one explanatory variable of each kind.

EXAMPLES

A comparison of incomes for females and males at a small college was done by comparing the mean salaries of the two groups. It was found that the mean for the men was larger than the mean for the women. However, it was pointed out that more women were in the lower ranks of the faculty, and perhaps that was

the reason for the disparity in salaries. To control for age, we could keep age constant and divide the data into groups, one for each age, and compare the salary means for men and women within each age group. With only a small data set to begin with, this would involve several analyses with very few cases.

Similarly, it would be possible to control for gender by doing a separate analysis for each gender. That would only involve two groups, but the groups would still be smaller than the entire faculty used as one group.

Instead, one multivariate regression analysis with age as one variable and gender as the dummy variable would achieve the same purpose without running into the problem with small data sets. The coefficient b_1 (the estimate of β_1) would show the effect of age, controlling for gender, and the coefficient b_2 (the estimate of β_2) would show the effect of gender, controlling for age.

It is also possible to do a more sophisticated analysis, allowing for the regression lines for income on age not to be parallel, which would allow for an interaction effect between age and gender on salary. This is done by creating an interaction variable as the product of the age variable and the dummy variable. This product variable is then used in a multiple regression analysis, with salary as the dependent variable and age, dummy, and interaction as three explanatory variables.

—Gudmund R. Iversen

REFERENCES

- Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics*, 13(3), 261–281.
- Hardy, M. A. (1993). *Regression with dummy variables* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–093). Newbury Park, CA: Sage.
- Wildt, A. R., & Ahtola, O. T. (1978). *Analysis of covariance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–012). Beverly Hills, CA: Sage.

ANALYSIS OF VARIANCE (ANOVA)

Analysis of variance is the name given to a collection of statistical methods used to analyze the impact of one or more NOMINAL VARIABLES as independent variables on a QUANTITATIVE VARIABLE as the dependent variable. For one nominal variable, there is ONE-WAY ANALYSIS OF VARIANCE; for two nominal variables,

there is TWO-WAY ANALYSIS OF VARIANCE and so on for higher models. Data on several dependent variables can be analyzed using MANOVA (MULTIVARIATE ANALYSIS OF VARIANCE).

EXPLANATION

A study of the impact of gender and race on income could employ analysis of variance to see whether mean incomes in the various gender/racial groups are different. It would perhaps be better to use a name such as *analysis of means* for these methods, but variances are actually used to determine whether group means are different—thus the name.

HISTORICAL ACCOUNT

Analysis of variance has its origins in the study of data from experiments. Many of the earliest applications were from agriculture in the study of how yield on a piece of land is affected by factors such as type of fertilizer and type of seed. The British statistician Sir Ronald A. Fisher did much of the early work developing analysis of variance.

Experiments can be set up in many different ways, and there are corresponding methods of analysis of variance for the analysis of the data from these designs. Thus, the name *analysis of variance* covers a large number of methods, and the ways in which an experiment is designed will determine which analyses are used. The ties between the design of EXPERIMENTS and analysis of variance are very strong.

There are also strong ties between analysis of variance and REGRESSION analysis. Historically, the two sets of statistical methods grew up separately, and there was little contact between the people who worked in the two areas. The old, elaborate methods of computations used in analysis of variance seemed very different from computations used in regression analysis. But now that statistical software has taken over the computations, it is not very difficult to demonstrate how the two sets of methods have many things in common. Most analyses of variance can be represented as regression analysis with DUMMY VARIABLES. Indeed, analysis of variance and regression analysis are special cases of the so-called general linear model. This became clearer when people started to use analysis of variance on data obtained not only from experiments but also from observational data.

APPLICATIONS AND EXAMPLES

One-Way Analysis of Variance

Many of the central ideas in analysis of variance can be illustrated using the simplest case in which there are data on a dependent variable Y from two groups. Thus, we have a nominal variable with two values as the independent variable. This becomes a one-way analysis of variance because there are data only on one independent variable. The data could have been obtained from a psychological experiment whereby one group consisted of the control group and the other group consisted of elements in an experimental group that received a certain stimulus. Then, is there any effect of the stimulus? Also, the data can come from an observational study in which there are incomes from people in different ethnic groups, and we want to see if the income means are different in the underlying ethnic populations.

Thus, we know what group an element (person) belongs to, and for each element, we have data on a quantitative variable Y (response, income). For the sake of numerical computing convenience from the days before electronic computing software, we use the notation commonly used in analysis of variance, which looks very different from the one used in regression analysis.

A typical observation of the variable Y is denoted y_{ij} . The first subscript i denotes the group the observation belongs to, and here $i = 1$ or $i = 2$. Within each group, the observations are numbered, and the number for a particular observation is denoted by j . For example, y_{25} is the observation for the fifth element in the second group. Similarly, the mean of the observations in group i is denoted $\bar{y}_i$, and the mean of all the observations is denoted by $\bar{y}$. With these symbols, we have the following simple identity:

$$y_{ij} - \bar{y} = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i). \quad (1)$$

The deviation of an observation y_{ij} from the overall mean $\bar{y}$ is written as a sum of the deviation of the group mean $\bar{y}_i$ from the overall mean $\bar{y}$ plus the deviation of the observation from the group mean.

If people's incomes were not affected by the group a person belongs to or by any other variables, then all the observations would be equal to the overall mean. Everyone would have the same income. There is absolutely no way for the observations to differ because they are not affected by any variables. The observed variations in the observations around the overall mean

$\bar{y}$ are then possibly due to two factors: the effect of the independent, categorical variable and the effect of all other variables.

Similarly, the first term on the right side measures how large the effect is of the independent variable. If this variable had no effect, then the means in the two groups would be the same, and the difference between a group mean and the overall mean would equal zero. The second term on the right measures how different an observation is from the group mean. If all other variables had no effects, then the observations in each group would all be equal because group membership is the only thing that matters. The differences between the observations and the group means are therefore due to the effects of all other variables, also known as the residual variable. In a population, this is known as the ERROR variable, and the estimated values in a sample form the residual variable:

$$\begin{array}{lcl} \text{Combined} & = & \text{Effect} \quad \quad + \text{Effect of} \\ \text{effect of} & & \text{of independent} \quad \text{residual variable} \\ \text{independent} & & \text{variable} \\ \text{and residual} & & \\ \text{variable} & & \end{array} \quad (2)$$

This identity holds for a single observation. Next, we want to summarize these effects across all the observations. We could simply add all the effects, but all three sums are equal to zero because the effects are all deviations from means.

Magnitudes of Effects

Instead, we square the left and right sides of the equation above to get positive terms. Squares are used for historical and mathematical convenience. There is nothing “correct” about using squares, and we have to realize that all the results we get from the analysis are heavily influenced by the choice of squares.

Adding the squares gives the sum of squared combined effects on the left side. On the right, we get the sum of squared effects for the independent variable, plus the sum of squared effects of the residual variable, plus two times the sum of the products of the effects of the independent and residual variables. It can then be shown that the sum of the cross-products always equals zero.

That gives the following equation for the squared effects across all the observations, written in symbols, words, and abbreviations:

$$\Sigma \Sigma (y_{ij} - \bar{y})^2 = \Sigma \Sigma (\bar{y}_i - \bar{y})^2 + \Sigma \Sigma (y_{ij} - \bar{y}_i)^2$$

$$\begin{array}{lcl} \text{Total sum of} & = & \text{Between-group} + \text{Within-group} \\ \text{squares} & & \text{sum of squares} \quad \text{squares} \end{array} \quad (3)$$

$$TSS = BSS + WSS \quad (4)$$

The two summations signs in front of the squares show that we add all the terms for the observations in each group and then add the terms for the two groups.

Thus, for each *individual*, we have broken down the combined effect into a sum of effects of the independent and residual variables. By using squares, we also have for the entire *group* of observations the combined effect broken down into the sum of effects of the independent variable (BETWEEN SUM OF SQUARES) and the residual variable. For more complicated analyses of variance, we get more extensive breakdowns of the total SUM OF SQUARES.

For the sake of simplicity, we can divide the three sums of squares by the total SUM OF SQUARES. That gives us the following:

$$\frac{TSS}{TSS} = \frac{BSS}{TSS} + \frac{WSS}{TSS}, \quad (6)$$

$$1.00 = R^2 + (1 - R^2). \quad (7)$$

This step defines the COEFFICIENT OF DETERMINATION R^2 . It tells how large the effect of the independent variable is as a proportion of the total sum of squares. This quantity is often denoted *eta*² in an older notation. It plays the same role as R^2 in regression analyses. R itself can be thought of as a correlation coefficient measuring the strength of the relationship between the independent and dependent variables.

The quantities described above are often displayed in an analysis of variance table, as shown in Table 1.

Hypothesis Testing

Could the relationship have occurred by chance? This question translates into the null hypothesis that the population group means are equal:

$$H_0: \mu_1 = \mu_2. \quad (8)$$

Table 1 Sums of Squares and Proportions

<i>Source of Effect</i>	<i>Sum of Squares</i>	<i>Proportion of Effect</i>
Independent variable	$\sum \sum (\bar{y}_i - \bar{y})^2$	R^2
Residual variable (error)	$\sum \sum (y_{ij} - \bar{y}_i)^2$	$1 - R^2$
Total	$\sum \sum (y_{ij} - \bar{y})^2$	1

With more than two groups and two population means, the test of the null hypothesis becomes more complicated. The null hypothesis states that all the population means are equal, but the logical opposite of several means being equal is only that at least *some* and not all of the means are different. After rejecting the null hypothesis, the question becomes *which* of the means are different.

A null hypothesis is studied by finding the corresponding *p* VALUE. Such a *p* value tells us how often (proportion of times) we get the data we got, or something more extreme, from a population in which the null hypothesis is true. This can also be seen as the probability of rejecting a true null hypothesis. If the *p* value is small—say, less than .05—we reject the null hypothesis. The *p* value is found from the data by computing the corresponding value of the statistical *F* variable. The actual computation is mostly of interest to statisticians only, but it is shown in Table 2, which corresponds to the table produced by most statistical software packages.

There are three more columns in Table 2 than in Table 1 that can be used to find the *p* value. The letter *k* stands for the number of groups, and *n* stands for the total number of observations in the study.

The computation of *F* is based on a comparison of the between and within sums of squares. But sums of squares cannot be compared directly because they depend on how many terms there are in the sums. Instead, each sum of squares is changed into a variance by dividing the sum by its appropriate DEGREES OF FREEDOM. That gives the quantities in the column labeled *Mean Square*, but the column could just as well have been denoted by the word *variance*. A mean square is an “average” square in the sense that a variance is an average squared distance from the mean. Finally, the ratio of the two mean squares is denoted by the letter *F* (in honor of Fisher).

By working through the mathematics of the derivations above, it is possible to show that when the

null hypothesis is true, then the two mean squares (variances) are about equal. Thus, the ratio of the two mean squares (the *F* value) is approximately equal to 1.00. When the null hypothesis is false and the population means are truly different, then the value of *F* is a good deal larger than 1.00.

When the values of the residual variable follow a normal distribution, the *F* variable has the corresponding *F*-DISTRIBUTION (a variable like the NORMAL DISTRIBUTION *t*, and CHI-SQUARE DISTRIBUTION variables). The *F*-distribution can then be used to find the *p* value for the data. The critical values of *F* are dependent on how many groups and how many observations there are. This is dealt with by saying that we have *F* with *k* − 1 and *n* − *k* degrees of freedom. Suppose we have data from two groups and a total of 100 observations, and *F* = 5.67. From the *F*-distribution with 2 − 1 = 1 and 100 − 2 = 98 degrees of freedom, we find that for this value of *F*, the *p* value = .02. Because it is small, we reject the null hypothesis and conclude that the difference between the two group means is statistically significant. Most statistical software provides an exact value of the *p* value. There are also tables for the *F* variable, but because there is a different *F*-distribution for each pair of degrees of freedom, the *F* tables tend to be large and not very detailed.

Data for two groups can also be analyzed using a *t*-test for the difference between two means or a regression analysis with a dummy variable. The (two-sided) *p* values would have been the same in all three cases. If we had more than two groups, there would not be any *t*-test, but we could use regression with the appropriate number of dummy variables instead of analysis of variance.

Two-Way Analysis of Variance

Here, two independent, nominal variables affect one dependent, quantitative variable. Similarly, there are three-way analysis or higher analyses, depending on how many independent variables there are.

Aside from having more than one independent variable, there is now a so-called interaction effect present in addition to the (main) effects of the independent variables. An interaction effect of the independent variables on the dependent variable is an effect of the two independent variables over and beyond their separate main effects.

Table 2 Analysis of Variance Table With Sums of Squares, Proportions, Mean Squares, and *F* Value

Source of Effect	Sum of Squares	Proportion of Effect	Degree of Freedom	Mean Square	<i>F</i> Value
Independent variable	$\sum \sum (\bar{y}_i - \bar{y})^2$ (<i>BSS</i>)	R^2	$k - 1$	$\frac{BSS}{k - 1} = BMS$	$\frac{BMS}{WMS}$
Residual variable (error)	$\sum \sum (y_{ij} - \bar{y}_i)^2$ (<i>WSS</i>)	$1 - R^2$	$n - k$	$\frac{WSS}{n - k} = WMS$	
Total	$\sum \sum (y_{ij} - \bar{y})^2$ (<i>TSS</i>)	1	$n - 1$		

The analysis is based on the following identity:

$$y_{ijk} - \bar{y} = (\bar{y}_i - \bar{y}) + (\bar{y}_{.j} - \bar{y}) + (\bar{y}_{ij} - \bar{y}_i - \bar{y}_{.j} + \bar{y}) + (\bar{y}_{ijk} - \bar{y}_{ij}). \quad (9)$$

Each observation in this notation is identified by the three subscripts *i*, *j*, and *k*. The first subscript gives the value of the first nominal variable, the second subscript gives the value of the second nominal variable, and the last subscript gives the number of observations for a particular combination of the two nominal variables.

If nothing influenced the dependent variable, all the values of that variable would have to be equal to the overall mean $\bar{y}$. Everyone would have the same income if there were no effects of gender, race, and all other variables. Any deviation from this value represents effects of one or more independent variables, as shown on the left side above. On the right side, the first difference represents the effect of the first nominal variable, the second difference represents the effect of the second nominal variable, the third term represents the effect of the interaction variable, and the last term is the residual, which measures the effects of all other variables. Without the inclusion of the interaction term, there would not be an equality.

By squaring the equation above, the right side gives the square of each term plus two times the cross-products. Adding across all the observations, the sums of all the cross-products equal zero, and the resulting sums of squares give the effects of the independent variables and the residual variable. These sums can be expressed as proportions, and by dividing by the residual mean square, we get *F* values and thereby *p* values for the tests of the hypotheses of no effects.

Models I, II, and III: Fixed- or Random-Effect Models

When we have data on the dependent variable for all values of the independent variables, we have a MODEL I ANOVA (fixed effects). When we have data on the dependent variable for only a sample of values of the independent variables, we have a MODEL II ANOVA (random effects). MODEL III ANOVA makes use of a mix of fixed and random effects.

—Gudmund R. Iversen

REFERENCES

- Bray, J. H., & Maxwell, S. E. (1985). *Multivariate analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-054). Beverly Hills, CA: Sage.
- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Iversen, G. R., & Norpoth, H. (1987). *Analysis of variance* (2nd ed., Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-001). Beverly Hills, CA: Sage.

ANALYTIC INDUCTION

Analytic induction is sometimes treated as a model for qualitative RESEARCH DESIGN. It portrays inquiry as an iterative process that goes through the following stages: initial definition of the phenomenon to be explained; examination of a few cases; formulation of a HYPOTHESIS; study of further cases to test this hypothesis; and, if the evidence negates the hypothesis, either redefinition of the phenomenon to exclude the negative case or reformulation of the hypothesis

so that it can explain it. This is followed by study of further cases, and if this further evidence supports the hypothesis, inquiry can be concluded. However, if any negative case arises in the future, inquiry will have to be resumed: Once again, either the hypothesis must be reformulated or the phenomenon redefined, and further cases must be studied, and so on.

Analytic induction shares with GROUNDED THEORY an opposition to the testing of hypotheses derived from armchair theorizing (see INDUCTION). It also shares with that approach an emphasis on the development of THEORY through investigating a relatively small number of cases and adopting a flexible mode of operation in which theory emerges out of—and guides future—data collection and analysis. At the same time, analytic induction has features that are at odds with the self-understanding of much QUALITATIVE RESEARCH today—notably, its explicit concern with testing hypotheses and discovering universal laws. Its most distinctive feature is recognition that the type of phenomenon to be explained may need to be reformulated in the course of research; this arises because the task is to discover theoretical categories all of whose instances are explained by a single type of cause. In the context of analytic induction, a theory specifies both necessary and sufficient conditions, and this is why a single exception is sufficient to refute any theory.

The term *analytic induction* seems to have been coined by the Polish American sociologist Florian Znaniecki in his book *The Method of Sociology* (Znaniecki, 1934). There, he contrasted it with *enumerative induction*, a term referring to the kind of inference that is characteristic of statistical method. Enumerative induction involves trying to discover the characteristics of phenomena belonging to a particular class by studying a large number of cases and describing what they have in common. Znaniecki points to a paradox here: If we do not know the essential features of a type of phenomenon, we cannot identify instances of it; however, if we *do* know those essential features, we already know what enumerative induction is designed to tell us. Following on from this, Znaniecki argued that, contrary to the claims of some contemporary social scientists, statistical method is not the method of natural science: Natural scientists do not study large numbers of cases and try to derive theories from their average characteristics. Rather, he suggested, they study a single case or a small number of cases in

depth, and develop theories that identify the essential features of these cases as instances of general classes of phenomena. Thus, according to Znaniecki, the task of science is the search for universal laws—relations of causality or functional dependence—not statistical generalizations; and it is analytic, not enumerative, induction that is its method (see Gomm, Hammersley, & Foster, 2000; Hammersley, 1989).

Znaniecki does not provide any detailed examples of analytic induction, but two subsequent writers provided what are now taken to be key illustrations. Alfred Lindesmith (1968) investigated the causal process involved in people becoming addicted to opiates, arguing that the essential condition was their coming consciously to use a drug to alleviate withdrawal symptoms. Donald Cressey, one of his students, set out to identify the necessary and sufficient conditions of trust violation, the latter concept being a theoretical reformulation of the commonsense and legal concept of embezzlement, a reformulation that emerged during the course of his research (Cressey, 1950, 1953). He put forward quite a complicated theory whereby those in positions of financial trust become trust violators when they have a financial problem that they cannot disclose to others; when they realize that this can be secretly solved through embezzlement; and when they can appeal to some account, an excuse, which preserves their sense of themselves as trustworthy.

There has been much controversy about the character and value of analytic induction. In an early but still valuable critique, W. P. Robinson (1951) challenged the distinction between enumerative and analytic induction; and argued that the latter, as usually practiced, focuses almost entirely on the discovery of necessary, rather than sufficient, conditions. He saw this as a weakness that could be overcome by researchers studying situations in which a theory would predict a particular type of phenomenon to occur. At the same time, he argued that requiring the evidence to display a universal rather than a probabilistic distribution—all instances in line with the theory and none at odds with it—is too demanding a standard because it is rarely likely to be met if a large number of cases are studied. Ralph Turner (1951) reacted against this argument by emphasizing the difference between analytic and enumerative induction, but proposed that they are complementary in framing explanations. Although the concept of analytic induction is less frequently invoked today by qualitative researchers than

previously, Howard Becker (1998) has recently reiterated and developed the idea that analytic induction best captures the logic of social scientific investigation.

—Martin Hammersley

REFERENCES

- Becker, H. S. (1998). *Tricks of the trade*. Chicago: University of Chicago Press.
- Cressey, D. (1950). The criminal violation of financial trust. *American Sociological Review*, 15, 738–743.
- Cressey, D. (1953). *Other people's money*. Glencoe, IL: Free Press.
- Gomm, R., Hammersley, M., & Foster, P. (Eds.). (2000). *Case study method*. London: Sage. [Includes the articles by Robinson and Turner.]
- Hammersley, M. (1989). *The dilemma of qualitative method*. London: Routledge Kegan Paul.
- Lindesmith, A. (1968). *Addiction and opiates*. Chicago: Aldine.
- Robinson, W. (1951). The logical structure of analytic induction. *American Sociological Review*, 16(6), 812–818.
- Turner, R. H. (1951). The quest for universals. *American Sociological Review*, 18(6), 604–611.
- Znaniecki, F. (1934). *The method of sociology*. New York: Farrar and Rinehart.

ANONYMITY

Anonymity refers to the assurance that participants are not individually identified by disguising or withholding their personal characteristics. In conducting social science research, the protection of “human subjects” is a serious concern for institutional review boards and researchers alike. Guaranteeing and preserving individuals’ anonymity is a key component of protecting their rights as research participants.

As researchers, we take several steps to ensure the anonymity of our participants. For example, we assign case numbers to each person instead of using respondents’ real names to identify their data. When presenting our findings, we disguise any facts that would allow participants to be identified. For example, we use pseudonyms instead of individuals’ real names. Researchers involved in the construction of data sets must remove all personal identifiers (such as name, address, and Social Security number) before permitting public use of the data.

We withhold direct identifiers as a matter of course, but indirect identifiers, such as state of residence,

organizational membership, and occupation, might also be used to recognize an individual or group (Inter-University Consortium for Political and Social Research, 2002). Quantitative researchers routinely recode the variables into broader categories as the broader categories adequately capture the shared characteristics they believe may affect the outcome variable of interest. Data from a sample of physicians, for example, may include the *type* of medical school from which they received their medical degrees (e.g., private or public institutions) rather than the *name* of that institution. In qualitative research, indirect identifiers may pose a greater problem, in part because data are not aggregated. Qualitative social scientists strive to provide rich narrative accounts of the group they are studying. Part of the richness of the narrative is the extent to which the author can link the participants’ social milieus to their experiences and perceptions. Thus, often we want the reader to know the occupation of a particular participant and how that may have affected other aspects of his or her life (or another variable of interest).

Protecting participants’ anonymity may be especially important when dealing with groups who are distrustful of research and science and/or those involved in illegal or stigmatized activities, such as illicit drug use. Groups who have been victimized or exploited by past research projects are likely to be distrustful. The Tuskegee Syphilis Study, in which African American men were left untreated so that the natural course of syphilis could be studied, is an egregious example.

In my research of women with HIV/AIDS, I discovered many reasons women worried about preserving their anonymity, including illicit drug use, criminal behavior of their children, and nondisclosure of their HIV/AIDS diagnosis to family and friends. Interestingly, on the other hand, some participants may not want to remain anonymous as they are proud—proud of what they have overcome, proud of what they have learned in the process, and proud of who they are or have become—and they want to share their stories. In fact, Grinyer (2002) has reported that a cancer patient who participated in her study expressed that she wanted her real name used rather than a pseudonym. The desire of some participants notwithstanding, researchers are obliged to take every possible measure to ensure the confidentiality of data and provide anonymity to research participants.

—Desirée Ciambrone

REFERENCES

- Grinyer, A. (2002). The anonymity of research participants: Assumptions, ethics, and practicalities. *Social Research Update*, 36, 1–4.
- Inter-University Consortium for Political and Social Research. (2002). *Guide to social science data preparation and archiving* [online]. Supported by the Robert Wood Johnson Foundation. Available: www.ICPSR.umich.edu

APPLIED QUALITATIVE RESEARCH

Applied qualitative research usually denotes qualitative studies undertaken explicitly to inform policymaking, especially in government but also in commerce. The term was possibly first coined in 1985 in Britain as part of a venture to encourage the use of qualitative research methods in government (Walker, 1985). At the time, public officials were inclined to equate evidence with OFFICIAL STATISTICS (Weiss, 1977), although qualitative methodologies had already successfully penetrated marketing and advertising research. Such studies are often commissioned by companies or government agencies and undertaken to a closely specified contract.

Applied qualitative research is an activity rather than a movement or coherent philosophy (Ritchie & Lewis, 2003). The objective is frequently to develop, monitor, or evaluate a policy or practice, using qualitative techniques as an alternative to, or to complement, other approaches. The rationale for using qualitative methods is that they are better suited to the task, either in specific circumstances or inherently. This may be because the topic is ill-defined or not well theorized, thereby precluding the development of structured questionnaires. It might be sensitive, intangible, or ephemeral, or it could appertain to distressing or emotional events or deep feelings. It could be inherently complex or an institution. Earlier research may have produced conflicting results.

Practitioners of applied qualitative research are therefore apt to align themselves with pragmatism, matching method to specific research questions. Their ontological position is likely to correspond to “subtle realism” (Hammersley, 1992), accepting that a diverse and multifaceted world exists independently of subjective understanding but believing it to be accessible via respondents’ interpretations. They will probably strive to be neutral and objective at each stage in the research

process and thereby generate findings that are valid and reliable, ones that are true to the beliefs and understandings of their respondents and potentially generalizable to other settings. They will seek detailed descriptions of the realities, as understood by their respondents, but clearly delineate between these interpretations and their own in analysis, making sense of complexity by simplification and structuring.

Applied qualitative research is eclectic in its use of qualitative methods although heavily reliant on IN-DEPTH INTERVIEWS and FOCUS GROUPS. There is currently a trend toward repeat interviews with the same respondents to investigate the links between intention, behavior, and outcomes. PROJECTIVE TECHNIQUES are commonly used, typically to stimulate respondents’ reflections on the process and outcome of projective exercises rather than to elicit scores on personality traits; occasionally, ETHNOGRAPHIC and ACTION RESEARCH techniques are also used. Different methods, including quantitative ones, are frequently used within the same project. When used with surveys, qualitative techniques are variously employed to refine concepts and develop questionnaires; to help interpret, illuminate, illustrate, and qualify survey findings; to triangulate between methods; and to provide complementary insight.

—Robert Walker

REFERENCES

- Hammersley, M. (1992). *What’s wrong with ethnography*. London: Routledge.
- Ritchie, J., & Lewis, J. (2003). *Qualitative research practice*. London: Sage.
- Walker, R. (Ed.). (1985). *Applied qualitative research*. Aldershot, UK: Gower.
- Weiss, C. (1977). *Use of social research in public policy*. Lexington, MA: D. C. Heath.

APPLIED RESEARCH

Applied research is, by definition, research that is conducted for practical reasons and thus often has an immediate application. The results are usually actionable and recommend changes to increase effectiveness and efficiency in selected areas. Applied research rarely seeks to advance theoretical frameworks. THEORY is generally used instrumentally to help

identify and define concepts and variables so that researchers can operationalize them for analysis.

Applied research is usually used in contrast to BASIC RESEARCH (sometimes called pure research) that does not yield immediate application upon completion. Basic research is designed to advance our understanding and knowledge of fundamental social processes and interactions, regardless of practical or immediate applications. This does not imply that applied research is more useful than basic research. Almost all basic research eventually results in some worthwhile application over the long run. The main distinction focuses on whether the research has *immediate use*. However, applied research often is easier to relate to by the general public as it yields actionable and practical results, whereas some basic research may have intuitive relevance only for researchers.

A common type of applied research is program evaluation. In some cases, researchers study the effectiveness of the current program to evaluate how effectively the program is in achieving its goal. In other cases, researchers might study the efficiency of the program's delivery system to examine the extent of resource wastage. In either case, the researchers' intent is to find ways to improve aspects of the program or its delivery system.

Another example of applied research is market research. Marketers may conduct research (focus groups, in-depth interviews, or surveys) to ferret consumers' product needs and preferences, enabling them to package the most appropriate products. Marketers may also test their communication materials with consumers to ensure that the intended marketing message is effectively delivered. The recommendations from this type of research could include word changes, packaging style, delivery approach, and even the color scheme of the materials.

In many cases, the audience of applied research is often limited to specific interested parties. As such, the results of applied research are often packaged as in-house reports that have limited circulation or are proprietary. However, some applied research is shared via publication in professional journals. For example, *The Gerontologist* has a section titled *Practice Concepts* that features applied research and short reports, in addition to publishing regular-length articles. An example from the *Practice Concepts* section is the report written by Si, Neufeld, and Dunbar (1999) on the relationship between having bedrails and residents' safety in a short-term nursing home rehabilitation unit.

Si et al. found that in the selected unit, the dramatic decline in bedrail use did not lead to an increase in residents' injuries. They suggested that the necessity of bedrails depended on both the training and experience of the staff as well as the type and severity of the patients' injuries.

For further reading, please refer to the following books: *Applied Research Design: A Practical Guide* by Hendrick, Bickman, and Rog (1993) and *Focus Groups: A Practical Guide for Applied Research* by Krueger (2000).

—VoonChin Phua

REFERENCES

- Hendrick, T., Bickman, L., & Rog, D. J. (1993). *Applied research design: A practical guide*. Newbury Park, CA: Sage.
- Krueger, R. A. (2000). *Focus groups: A practical guide for applied research*. Thousand Oaks, CA: Sage.
- Si, M., Neufeld, R. R., & Dunbar, J. (1999). Removal of bedrails on a short-term nursing home rehabilitation unit. *The Gerontologist*, 39, 611–614.

ARCHIVAL RESEARCH

For the social scientist, archival research can be defined as the locating, evaluating, and systematic interpretation and analysis of sources found in archives.

Original source materials may be consulted and analyzed for purposes other than those for which they were originally collected—to ask new questions of old data, provide a comparison over time or between geographic areas, verify or challenge existing findings, or draw together evidence from disparate sources to provide a bigger picture.

For the social scientist, using archives may seem relatively unexciting compared to undertaking fieldwork, which is seen as fresh, vibrant, and essential for many researchers. Historical research methods and archival techniques may seem unfamiliar; for example, social scientists are not used to working with handbreak written sources that historians routinely consult. However, these methods should be viewed as complementary, not competing, and their advantages should be recognized for addressing certain research questions. Indeed, consulting archival sources enables the social scientist to both enhance and challenge the established methods of defining and collecting data.

What, then, are the types of sources to be found of interest to the social scientist? Archives contain official sources (such as government papers), organizational records, medical records, personal collections, and other contextual materials. Primary research data are also found, such as those created by the investigator during the research process, and include transcripts or tapes of interviews, FIELDNOTES, personal DIARIES, observations, unpublished manuscripts, and associated correspondence. Many university repositories or special collections contain rich stocks of these materials. Archives also comprise the cultural and material residues of both institutional and theoretical or intellectual processes, for example, in the development of ideas within a key social science department.

However, archives are necessarily a product of sedimentation over the years—collections may be subject to erosion or fragmentation—by natural (e.g., accidental loss or damage) or manmade causes (e.g., purposive selection or organizational disposal policies). Material is therefore subjectively judged to be worthy of preservation by either the depositor or archivist and therefore may not represent the original collection in its entirety. Storage space may also have had a big impact on what was initially acquired or kept from a collection that was offered to an archive.

Techniques for locating archival material are akin to excavation—what you can analyze depends on what you can find. Once located, a researcher is free to immerse himself or herself in the materials—to evaluate, review, and reclassify data; test out prior hypotheses; or discover emerging issues. Evaluating sources to establish their validity and reliability is an essential first step in preparing and analyzing archival data. A number of traditional social science analytic approaches can be used once the data sources are selected and assembled—for example, GROUNDED THEORY to uncover patterns and themes in the data, BIOGRAPHICAL METHODS to document lives and ideas by verifying personal documentary sources, or CONTENT ANALYSIS to determine the presence of certain words or concepts within text.

Access to archives should not be taken for granted, and the researcher must typically make a prior appointment to consult materials and show proof of status. For some collections, specific approval must be gained in advance via the archivist.

—Louise Corti

REFERENCES

- Elder, G., Pavalko, E. K., & Clipp, E. C. (1993). *Working with archival lives: Studying lives* (Sage University Papers on Quantitative Applications in the Social Sciences, 07–088). Newbury Park, CA: Sage.
- Foster, J., & Sheppard, J. (1995). *British archives: A guide to archival resources in the United Kingdom* (3rd ed.). London: Macmillan.
- Hill, M. M. (1993). *Archival strategies and techniques* (Qualitative Research Methods Series No. 31). Newbury Park, CA: Sage.
- Scott, J. (1990). *A matter of record: Documentary sources in social research*. Cambridge, UK: Polity.

ARCHIVING QUALITATIVE DATA

Qualitative data archiving is the long-term preservation of qualitative data in a format that can be accessed by researchers, now and in the future. The key to ensuring long-term accessibility lies in the strategies for data processing, the creation of informative documentation, and in the physical and technical procedures for storage, preservation, security, and access (see DATA ARCHIVES).

QUALITATIVE DATA are data collected using QUALITATIVE RESEARCH methodology and techniques across the range of social science disciplines. Qualitative research often encompasses a diversity of methods and tools rather than a single one, and the types of data collected depend on the aim of the study, the nature of the sample, and the discipline. As a result, data types extend to IN-DEPTH or UNSTRUCTURED INTERVIEWS, SEMI-STRUCTURED INTERVIEWS, FIELDNOTES, unstructured DIARIES, observations, personal documents, and photographs. Qualitative research often involves producing large amounts of raw data, although the methods typically employ small sample sizes. Finally, these data may be created in a number of different formats: digital, paper (typed and handwritten), and audiovisual.

Until 1994, no procedures existed for the systematic archiving and dissemination of qualitative data, although the ORAL HISTORY community had a professional interest in preserving tape recordings gathered from oral history interviewing projects. Many of these tape recordings may be found at the British Library National Sound Archive in London.

In 1994, the first qualitative data-archiving project on a national scale was established in the United

Kingdom, with support from the Economic and Social Research Council. The Qualidata Centre established procedures for sorting, describing, processing both raw data and accompanying documentation (meta-data), and establishing appropriate mechanisms for access.

The way data sets are processed can be split into three main activities: checking and anonymizing, converting the data set, and generating meta-data. Checking the data set consists of carrying out activities such as establishing the completeness and quality of data, as well as the relationships between data items (e.g., interviews, field notes, and recordings). Good-quality transcription is essential when archiving interviews without audio recordings, as is ensuring that the content of the data meets any prior consent agreements with participants. CONFIDENTIALITY and copyright are two key issues that arise in the archiving and dissemination of qualitative data. Converting data consists of transferring data to a format suitable for both preservation and dissemination. This may include digitization of paper resources.

The third main processing activity, generating meta-data (data about data), refers to the contextual information generated during processing, such as the creation of a "data list" of interview details, and ensuring that speaker and interviewer tags and questions or topic guide headers are added to raw text. A key function of meta-data is to enable users to locate transcripts or specific items in a data collection most relevant to their research. User guides are also created that bring together key documentation such as topic guides, personal research diaries, end of award reports, and publications. This type of information, collated by principal investigators at the fieldwork and analysis stage, is clearly of great benefit for future archiving.

Finally, the reuse or SECONDARY ANALYSIS OF QUALITATIVE DATA is a topic that has begun to spark debate since the late 1990s. Qualitative data can be approached by a new analyst in much the same way as quantitative data can—to ask new questions of old data, provide comparisons, or verify or challenge existing findings. However, the shortcomings of using "old data" for secondary analysis should always be recognized.

First, the lack of the exact data required to answer the question at hand can be a frustrating experience. Second, as Hammersley (1997) notes, an assessment of the quality and validity of the data without access to full knowledge of the research process is more complicated than appraising one's own data collection. This is because the path of qualitative analysis is almost never linear and is almost certainly likely to involve a degree

of trial and error in the pursuit of interesting lines of investigation.

Third, for qualitative data in particular, there is the difficulty of reconstructing the interview situation or fieldwork experience, which is often not fully captured in an archived interview transcript or set of field notes. Consequently, a secondary analyst does not have access to the tacit knowledge gained by the primary researcher, which, in some circumstances, may be important in helping interpret data. Listening to archived audio-recorded interviews, in addition to reading transcripts, can often help to reconstruct an interview setting.

—Louise Corti

REFERENCES

- Corti, L., Foster, J., & Thompson, P. (1995). *Archiving qualitative research data* (Social Research Update No. 10). Surrey, UK: Department of Sociology, University of Surrey.
- Hammersley, M. (1997). Qualitative data archiving: Some reflections on its prospects and problems. *Sociology*, 31(1), 131–142.
- James, J. B., & Sørensen, A. (2000, December). Archiving longitudinal data for future research: Why qualitative data add to a study's usefulness. *Forum Qualitative Sozialforschung* [Forum: Qualitative Social Research] [online journal], 1(3). Available: <http://qualitative-research.net/fqs/fqs-eng.htm>
- Mruck, K., Corti, L., Kluge, S., & Opitz, D. (Eds.). (2000, December). Text archive: Re-analysis. *Forum Qualitative Sozialforschung* [Forum: Qualitative Social Research] [online journal], 1(3). Available: <http://qualitative-research.net/fqs/fqs-eng.htm>

ARIMA

ARIMA is an acronym for autoregressive, integrated, moving average, denoting the three components of a general class of STOCHASTIC time-series models described by Box and Tiao (1965, 1975) and developed by Box and Jenkins (1976). Estimated univariate ARIMA models are often used successfully for forecasting and also form the basic foundation for multivariate analysis, such as INTERVENTION ANALYSIS assessment or TRANSFER FUNCTION models in the BOX-JENKINS MODELING approach to TIME SERIES. The use of ARIMA models in conjunction with transfer function models is often contrasted with the use of time-series REGRESSION models, but ARIMA models are increasingly used in tandem with regression models as well.

THE PIECES

ARIMA models have three components, with each one describing a key feature of a given time-series process. Briefly, the three components correspond to the effect of past values, called an autoregressive (AR) process; past shocks, called a moving average (MA) process; and the presence of a stochastic trend, called an integrated (I) process. Each of these is discussed below with reference to example time-series processes.

Autoregressive

Time-series processes are often influenced by their own past values. Consider the percentage of the public that says they approve of the way the president is doing his job. Public sentiment about the president's performance today depends in part on how they viewed his performance in the recent past. More generally, a time series is autoregressive if the current value of the process is a function of past values of the process. The effect of past values of the process is represented by the autoregressive portion of the ARIMA model, written generally as follows:

$$y_t = a + b_1 y_{t-1} + b_2 y_{t-2} + \cdots + b_p y_{t-p} + \mu_t. \quad (1)$$

Each observation consists of a random shock and a linear combination of prior observations. The number of lagged values determines the order of autoregression and is given by p , denoted $AR(p)$.

Moving Average

ARIMA models assume that time-series processes are driven by a series of random shocks. Random shocks may be loosely thought of as a collection of inputs occurring randomly over time. More technically, random shocks are normally and identically distributed random variables with mean zero and constant variance. As an example, consider that reporting requirements for violent crime may change. This change "shocks" statistics about violent crime so that today's reports are a function of this (and other) shock(s). We can write this as follows:

$$y_t = a + \mu_t + c_1 \mu_{t-1} + c_2 \mu_{t-2} + \cdots + c_q \mu_{t-q}, \quad (2)$$

where q gives the number of past shocks, μ that affect the current value of the $\bar{y}$ process, here violent crime statistics. The number of lagged values of μ determines the moving average *order* and is given by q , denoted $MA(q)$.

Integrated

ARIMA models assume that the effects of past shocks and past values decay over time: Current values of presidential approval reflect recent shocks and recent values of approval more so than distant values. For example, mistakes a president makes early in his tenure are forgotten over time. This implies that the `PARAMETER` values in the general AR and MA model components must be restricted to ensure that effects decay (i.e., that the time series is stationary). Loosely, a stationary time series is one with constant mean, variance, and covariances over time. A series that is not stationary is known as *integrated*. Some economic time series (e.g., economic growth) are thought to be integrated because they exhibit a historical tendency to grow.

Integrated time series must be transformed before analysis can take place. The most common `TRANSFORMATION` is called differencing, whereby we analyze the changes in y rather than the levels of y (we analyze $y_t - y_{t-1}$). The number of times a series must be differenced to be stationary gives the order of integration, denoted $I(d)$, where d may be any real number (if d is a fraction, the process is said to be fractionally integrated, and the model is called an ARFIMA model). If the time series is integrated in levels but is not integrated in changes, the process is integrated of order $d = 1$.

The condition for stationarity for an $AR(1)$ process is that b_1 must be between -1 and $+1$. If b_1 is not in this range, the effects of the past values of the process would accumulate over time, and the values of successive y s would move toward infinity; that is, the series would not be stationary and would need to be differenced before we (re)consider the order of autoregression that characterizes the process. In the $MA(1)$ case, the restrictions require c_1 to be between -1 and $+1$. In the case of higher order processes, similar restrictions must be placed on the parameters to ensure stationarity.

THE GENERAL MODEL

The general ARIMA(p, d, q) model combines the model components given above and is written as follows:

$$y_t = a + b_1 y_{t-1} + b_2 y_{t-2} + \cdots + b_p y_{t-p} + \mu_t + c_1 \mu_{t-1} + c_2 \mu_{t-2} + \cdots + c_q \mu_{t-q}, \quad (3)$$

where y has been differenced a suitable number (d , generally 0 or 1) of times to produce stationarity, p gives the number of AR lags, and q is the number of MA lags. Any time series may contain any number (including 0) of AR, I, and MA parameters, but the number of parameters will tend to be small. Generally, a simple structure underlies time-series processes; most processes are well explained as a function of just a few past shocks and past values. In fact, many social science processes behave as simple ARIMA(1,0,0) processes.

It is important to note that a given time-series process may be best characterized as an ARIMA(0,0,0) process, also called a WHITE-NOISE process. A white-noise process is a series of random shocks. It does not have autoregressive, integrated, or moving average components. When estimating regression models, inferences are only valid if the error is a white-noise process.

The ARIMA model that characterizes a particular series is identified through a specific methodology. In particular, an ARIMA model of the process that we care about is built from an observed time series in an empirically driven process of identification, estimation, and diagnosis (see BOX-JENKINS MODELING for an explanation of how this is done). For a more general treatment of ARIMA models and the Box-Jenkins approach, see McCleary and Hay (1980).

—Suzanna De Boef

REFERENCES

- Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.
- Box, G. E. P., & Tiao, G. C. (1965). A change in level of nonstationary time series. *Biometrika*, 52, 181–192.
- Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70, 70–79.
- McCleary, R., & Hay, R. A., Jr. (1980). *Applied time series analysis for the social sciences*. London: Sage.

ARROW'S IMPOSSIBILITY THEOREM

Kenneth J. Arrow (1921–) is one of the 20th century's leading economists. His most famous contribution to the social sciences is Arrow's impossibility theorem (Arrow, 1963). Arrow examined the logic of choice by a society based on the preferences of the individuals in that society and identified several reasonable conditions that a social choice rule used by a society should have. The impossibility theorem states that when there are three or more alternatives, it is impossible to find a social choice rule that always aggregates individual preferences into social preferences in a way that satisfies these desirable conditions. Thus, even if each individual's preferences are perfectly consistent and logical, the preferences of society as a whole might be incoherent.

Begin by assuming each individual in a society has preferences over a list of alternatives. A social choice rule is a rule that takes all individual preferences and combines them to determine the preferences of society as a whole. For example, individuals might have preferences over political candidates, and we seek a social choice rule (such as majority rule) to combine their preferences and determine who should take office. Arrow identified five reasonable conditions that a social choice rule should satisfy. The social choice rule should be the following:

- *Complete*. The social choice rule should provide a complete ranking of all alternatives. That is, for any alternatives A and B, the social choice rule should tell us if A is preferred to B, B is preferred to A, or if there is social indifference between A and B.
- *Paretian*. If every individual prefers A to B, then the social choice rule should rank A above B. Selecting alternative A over alternative B in this case would be a Pareto-optimal move (one that makes all individuals better off), and the social choice rule should rank A above B.
- *Transitive*. If the social choice rule ranks A above B and B above C, then A should be ranked higher than C.
- *Independent of irrelevant alternatives*. The ranking of A compared to B should not depend on preferences for other alternatives.

- *Nondictatorial*. The social choice rule should not depend on the preferences of only one individual (a dictator).

The proof of Arrow's impossibility theorem demonstrates that it is impossible to find a social choice rule that always satisfies all five of these conditions when there are three or more alternatives. That is, no matter what rule a society adopts to determine social preferences, it will either have some undesirable "irrational" features or be determined entirely by one person.

The result of the impossibility theorem is troubling to those who study democratic choice. However, relaxing some of the conditions imposed by the impossibility theorem or adding additional restrictions does allow for making social decisions through the political process. An important example is when alternatives are assumed to be arranged along some one-dimensional spectrum, and individual preferences over these alternatives are single-peaked, meaning that strength of preference peaks at the most preferred alternative and decreases across other alternatives as they become further from the most preferred alternative on the spectrum. In this case, with an odd number of individuals, majority voting will lead to adopting the preferences of the median voter (e.g., Black, 1958; Downs, 1957). Societal norms and institutions can also offer a means to escape the implications of the impossibility theorem (e.g., Riker, 1980).

—Garrett Glasgow

REFERENCES

- Arrow, K. J. (1963). *Social choice and individual values* (2nd ed.). New Haven, CT: Yale University Press.
- Black, D. (1958). *The theory of committees and elections*. Cambridge, UK: Cambridge University Press.
- Downs, A. (1957). *An economic theory of democracy*. New York: Harper Collins.
- Riker, W. H. (1980). Implications from the disequilibrium of majority rule for the study of institutions. *American Political Science Review*, 74, 432–446.

ARTIFACTS IN RESEARCH PROCESS

The term *research artifacts* refers to the systematic biases, uncontrolled and unintentional, that can threaten the INTERNAL or EXTERNAL VALIDITY of one's research conclusions. The potential for research artifacts in social science research exists because human participants are sentient, active organisms rather than

passive, inactive objects. As such, the interactions between the researcher and the agents of study can potentially alter the experimental situation in subtle ways unintended by the researcher.

The potential problems that research artifacts may cause in social science research were evident in the late 19th and early 20th centuries. For example, a controversial series of studies conducted at the Western Electric Company's Hawthorne Works in Illinois during the 1920s was interpreted as implying that simply observing someone's behavior is sufficient to alter that person's behavior. Earlier, the work of the eminent German psychologist Oskar Pfungst (1911/1965), involving a horse named "Clever Hans," at the turn of the century dramatically demonstrated how an individual's expectations could unconsciously influence another's behavior. In another classic work, Saul Rosenzweig (1933) argued that the experimental situation should be considered a psychological problem in its own right. He suggested that one must consider the possibility that the attitudes and motivations of the experimenter and research participant can subtly influence the research situation. Despite these warning signs of the potential problem of research artifacts, it was not until the late 1950s and 1960s that this problem and its ramifications for social science research were systematically investigated. This research, which was first pulled together in a volume edited by Rosenthal and Rosnow (1969), has commonly focused on two potential sources of research artifacts: the researcher and the research participant.

RESEARCHER-RELATED ARTIFACTS

Because researchers are not passive, disinterested participants in the scientific process, they can sometimes unknowingly or unwittingly engage in behaviors that introduce an artifact into the research design. Robert Rosenthal (1966), in a seminal book, defined two broad types of researcher-related artifacts as noninteractional and interactional artifacts. Noninteractional artifacts are biases that do not actually affect the research participants' behavior. Rather, they are systematic biases in how the researcher observes, interprets, or reports the research results. For example, the observations the researcher makes may be biased in a manner that increases the likelihood that the researcher's hypotheses will be supported. Similarly, how the researcher interprets the data may be biased in the direction of supporting the hypotheses of interest.

Although these two types of noninteractional biases may be unintentional, a third type is clearly intentional. The researcher may fabricate or fraudulently manipulate data to manufacture support for his or her hypotheses, which is clearly unethical.

The second category of researcher-related artifacts is associated with the interaction between the researcher and the research participants. For example, various biosocial characteristics of the researcher (such as ethnicity, sex, and age) have the potential to influence the participants' behavior in an unintended manner. Likewise, psychosocial attributes of the researcher (such as personality) may lead participants to respond in a manner ostensibly unrelated to the factors under investigation. Situational effects (such as changes in the researcher's behavior during the course of the study) can alter the manner in which individuals respond to the experimental stimuli. Finally, as Rosenthal (1966) demonstrated, even the simple knowledge of one's hypotheses can influence the researcher's behavior, leading to a self-fulfilling prophecy or EXPERIMENTER EXPECTANCY bias, which is by far the most thoroughly investigated source of interactional artifacts.

PARTICIPANT ARTIFACTS

The research participants are also active, motivated parties who are cognizant of their role as the object of study. Martin T. Orne (1962) argued that VOLUNTEER SUBJECTS, in particular, are motivated by a desire to help the cause of science and therefore will adopt the role of a "good subject." In this capacity, they seek out and respond to cues (DEMAND CHARACTERISTICS) that ostensibly enable them to behave in ways that are likely to support the researcher's hypotheses. Milton J. Rosenberg (1969), however, argued that EVALUATION APPREHENSION motivates research participants' behaviors. Their concern over being negatively evaluated, he argued, leads them to respond to experimental stimuli in a manner that presumably enables them to "look good." A third possible motivation underlying the participants' behavior in the research setting is the desire to obey the investigator who is perceived as an authority figure. Regardless of the underlying motivation, the implication is that individuals are responding to the stimuli under investigation in a systematically biased fashion.

MINIMIZING RESEARCH ARTIFACTS

Work on identifying potential sources of research artifacts has implied ways to either detect the presence of an artifact or minimize its introduction into the RESEARCH DESIGN. For example, potential noninteractional research-related artifacts may be identified through independent REPLICATION of research results and unimpeded access to research data by other scientists. Rosenthal (1966) described how an expectancy control design could be used to detect the presence of experimenter expectancy bias. Standardized research protocols, DOUBLE-BLIND PROCEDURE experimental designs, and the separation in time and place of the experimental treatment and measurement of the treatment effects represent other general strategies for minimizing researcher-related artifacts.

Strategies for avoiding certain participant artifacts include the introduction of alternative task-orienting cues and the creation of an experimental situation that reduces evaluation apprehension concerns or the motivation to respond to demand cues. Using participants who are unaware that they are participating in a research study can minimize the REACTIVITY problem, but this strategy can raise ethical issues regarding the invasion of PRIVACY.

Orne (1969) described how quasi-control subjects could be used to ferret out demand characteristics operating in a study. These participants essentially act as "consultants" to the researcher. Quasi-control subjects are asked to discuss the rationale behind their behavior after exposure to various aspects of the experimental design or speculate on how they might respond if they were actually exposed to the experimental stimuli. In the end, the most important mechanism that researchers have for identifying and controlling potential research artifacts is the independent replication of research results by different researchers.

—David B. Strohmetz and Ralph L. Rosnow

REFERENCES

- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776–783.
- Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 143–179). New York: Academic Press.

- Pfungst, O. (1965). *Clever Hans: The horse of Mr. von Osten*. New York: Holt, Rinehart, & Winston. (Original work published 1911.)
- Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 279–349). New York: Academic Press.
- Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.
- Rosenthal, R., & Rosnow, R. L. (Eds.). (1969). *Artifact in behavioral research*. New York: Academic Press.
- Rosenzweig, S. (1933). The experimental situation as a psychological problem. *Psychological Review*, 40, 337–354.
- Rosnow, R. L., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. New York: W. H. Freeman.
- Strohmets, D. B., & Rosnow, R. L. (1994). A mediational model of research artifacts. In J. Brzezinski (Ed.), *Probability in theory-building: Experimental and nonexperimental approaches to scientific research in psychology* (pp. 177–196). Amsterdam: Editions Rodopi.

ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) is the study of how to build intelligent systems. Specifically, AI is concerned with developing computer programs with intelligent behaviors, such as problem solving, reasoning, and learning.

Since the name *artificial intelligence* was coined in 1956, AI has been through cycles of ups and downs. From the beginning, game playing has been an important AI research subject for problem solving. In 1997, IBM's Deep Blue chess system defeated world-champion Garry Kasparov. Expert systems have been successfully developed to achieve expert-level performance in real-world applications by capturing domain-specific knowledge from human experts in knowledge bases for machine reasoning, and they have been used widely in science, medicine, business, and education. Machine learning arose as an important research subject in AI, enabling dynamic intelligent systems with learning ability to be built. Patterned after biological evolution, genetic algorithm spawns the population of competing candidate solutions and drives them to evolve ever better solutions. Agent-based models demonstrate that globally intelligent behavior can arise from the interaction of relatively simple structures. Based on the interaction between individual agents, intelligence is seen as emerging from

society and not just as a property of an individual agent. Recently, AI has been moving toward building intelligent agents in real environments, such as Internet auctions. Multiagent systems, as the platform for the convergence of various AI technologies, benefit from the abstractions of human society in an environment that contains multiple agents with different capabilities, goals, and beliefs.

The AI approach has been applied to social, psychological, cognitive, and language phenomena. At the theory-building level, AI techniques have been used for theory formalization and simulation of societies and organizations. For example, a GAME THEORY problem, the "prisoner's dilemma," is a classic research problem that benefits from computer SIMULATION in exploring the conditions under which cooperation may arise between self-interested actors who are motivated to violate agreements to cooperate at least in the short term. More recently, the multiagent system emerged as a new approach of modeling social life. These models show how patterns of diffusion of information, emergence of norms, coordination of conventions, and participation in collective action can be seen as emerging from interactions among adaptive agents who influence each other in response to the influence they receive.

At a data analysis level, AI techniques have been used to intelligently analyze qualitative data (data collected using methods such as ethnography) using symbolic processors and quantitative data using AI-enhanced statistical procedures. For example, it has been shown that NEURAL NETWORKS can readily substitute for and even outperform multiple regression and other multivariate techniques. More recently, the process of knowledge discovery has shown great potential in social science research by taking the results from data mining (the process of extracting trends or patterns from data) and transforming them into useful and interpretable knowledge through the use of AI techniques.

—Xiaoling Shu

REFERENCES

- Alonso, E. (2002). AI and agents: State of the art. *AI Magazine*, 23(3), 25–29.
- Bainbridge, E. E., Carley, K. M., Heise, D. R., Macy, M. W., Markovsky, B., & Skvoretz, J. (1994). Artificial social intelligence. *Annual Review of Sociology*, 20, 407–436.
- Conte, R., & Dellarocas, C. (2001). Social order in info societies: An old challenge for innovation. In R. Conte &

- C. Dellarocas (Eds.), *Social order in multi-agent systems* (pp. 1–16). Boston: Kluwer Academic.
- Luger, G. F. (2002). *Artificial intelligence: Structures and strategies for complex problem solving* (4th ed.). Reading, MA: Addison-Wesley.
- Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, 28, 143–166.

ARTIFICIAL NEURAL NETWORK. See NEURAL NETWORK

ASSOCIATION

The analysis of the relation (association) between variables is a fundamental task of social research. The association between two (categorical) variables can be expressed in a two-way CONTINGENCY TABLE or in cross-classification by calculating row percentages, column percentages, or total percentages. This technique can also be implemented if variables are continuous. The variables must be grouped into categories. A summary measure of the association between two variables is given by association coefficients, which measure the strength of the relation between two variables X and Y and, for ordinal, interval, and ratio variables, the direction of the association. The direction can be positive (concordant; higher values in X correspond to higher values in Y) or negative (discordant; higher values in X correspond to lower values in Y).

Two broad classes of association coefficients can be distinguished:

- **ASYMMETRIC** (or directional) **MEASURES** or association coefficients assume that one of the two variables (e.g., X) can be identified as the independent variable and the other variable (e.g., Y) as the dependent variable. For example, the variable X might be gender, and variable Y might be occupational status.
- **SYMMETRIC MEASURES** or association coefficients do not require differentiation between dependent and independent variables. They can, therefore, be used if such a distinction is impossible; an example is a researcher who is interested in the relations between different leisure-time activities (doing

sports and visiting the theater). However, symmetric coefficients can also be computed if identification as independent and dependent variables is possible.

Some examples of symmetric coefficients (e.g., Healey, 1995; SPSS, 2002) are the following: ϕ (ϕ); Cramér's V ; the contingency coefficient C for nominal variables; Kendall's τ_a (τ_a), τ_b (τ_b), and τ_c (τ_c); Goodman and Kruskal's gamma (γ) for ordinal variables; Spearman's rho (ρ) for rank (ordinal) data; and Pearson's r for interval or ratio variables. Both the size of the table and the number of both variables' categories determine which measure is used within one measurement level. Directed symmetric coefficients are often synonymously labeled as **CORRELATION** coefficients; sometimes, the term *correlation* is reserved for directed symmetric associations between two continuous variables.

Examples of asymmetric coefficients (e.g., Healey, 1995; SPSS, 2002) are lambda (λ), Goodman and Kruskal's τ , the uncertainty coefficient for nominal variables, Somers's d for ordinal variables, and eta (η), if the dependent variable is interval or ratio scaled and the independent variable is nominal. Further symmetric coefficients, such as symmetric lambda, symmetric Somers, and so on, can be derived from asymmetric coefficients by means of averaging.

The quoted coefficients (except η) assume that both variables have the same measurement level. If the analyzed variables have different measurement levels, it is recommended to reduce the measurement level and then compute an appropriate association measure for the lower measurement level. Unfortunately, this is often necessary as coefficients for mixed measurement levels are not available in standard statistical software packages. Coefficients for mixed measurement levels can be derived from Daniels' generalized coefficient of correlation, for example (Daniels, 1944; Kendall, 1962, pp. 19–33).

The concept of association can also be generally applied to the multivariate case. Examples of multivariate association coefficients include the **CANONICAL CORRELATION ANALYSIS** between two sets of variables, $X = \{X_1, X_2, \dots, X_p\}$ and $Y = \{Y_1, Y_2, \dots, Y_p\}$; the **MULTIPLE CORRELATION** coefficient R (from multiple regression) for a set of variables, $X = \{X_1, X_2, \dots, X_p\}$ and $Y = \{Y_1\}$; and the inertia between two nominal variables, X and Y , from **BIVARIATE ANALYSIS**.

An alternative strategy to analyze bivariate and multivariate contingency tables is ASSOCIATION MODELS based on LOG-LINEAR MODELING. Log-linear models allow one to specify and analyze different and complex patterns of associations, whereas association coefficients give a summary measure.

Association coefficients measure the strength of the relation between two variables. Users oftentimes demand guidelines to interpret the strength, but general guidelines are impossible. The strength of an association depends on the sample's homogeneity, the reliability of the variables, and the type of relationship between the variables. In psychological experiments (very homogeneous sample, variables with high reliability, direct causal link), a coefficient of 0.4 may signal weak correlation, whereas the same value could be suspiciously high in a sociological survey (heterogeneous sample, lower reliability, indirect causal links). Hence, when fixing threshold values, it is important to consider these factors along with the results of previous studies.

—Johann Bacher

REFERENCES

- Daniels, H. E. (1944). The relation between measures of correlation in the universe of sample permutations. *Biometrika*, 35, 129–135.
- Healey, J. F. (1995). *Statistics: A tool for social research* (3rd ed.). Belmont, CA: Wadsworth.
- Kendall, M. G. (1962). *Rank correlation methods* (3rd ed.). London: Griffin.
- SPSS. (2002, August). Crosstabs. In *SPSS 11.0 statistical algorithms* [online]. Available: <http://www.spss.com/tech/stat/algorithms/11.0/crosstabs.pdf>

ASSOCIATION MODEL

Although the term ASSOCIATION is used broadly, *association model* has a specific meaning in the literature on CATEGORICAL DATA ANALYSIS. By *association model*, we refer to a class of statistical models that fit observed frequencies in a cross-classified table with the objective of measuring the strength of association between two or more ordered categorical variables. For a two-way table, the strength of association being measured is between the two categorical variables that comprise the cross-classified table. For three-way or

higher-way tables, the strength of association being measured can be between any pair of ordered categorical variables that comprise the cross-classified table. Although some association models make use of the a priori ordering of the categories, other models do not begin with such an assumption and indeed reveal the ordering of the categories through estimation. The association model is a special case of a LOG-LINEAR MODEL or log-bilinear model.

Leo Goodman should be given credit for having developed association models. His 1979 paper, published in the *Journal of the American Statistical Association*, set the foundation for the field. This seminal paper was included along with other relevant papers in his 1984 book, *The Analysis of Cross-Classified Data Having Ordered Categories*. Here I first present the canonical case for a two-way table before discussing extensions for three-way and higher-way tables. I will also give three examples in sociology and demography to illustrate the usefulness of association models.

GENERAL SETUP FOR A TWO-WAY CROSS-CLASSIFIED TABLE

For the cell of the i th row and the j th column ($i = 1, \dots, I$, and $j = 1, \dots, J$) in a two-way table of R and C , let f_{ij} denote the observed frequency and F_{ij} the expected frequency under some model. Without loss of generality, a log-linear model for the table can be written as follows:

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + \mu_{ij}^{RC}, \quad (1)$$

where μ is the main effect, μ_i^R is the row effect, μ_j^C is the column effect, and μ_{ij}^{RC} is the interaction effect on the logarithm of the expected frequency. All the parameters in equation (1) are subject to ANOVA-type normalization constraints (see Powers & Xie, 2000, pp. 108–110). It is common to leave μ_i^R and μ_j^C unconstrained and estimated nonparametrically. This practice is also called the “saturation” of the marginal distributions of the row and column variables. What is of special interest is μ_{ij}^{RC} : At one extreme, μ_{ij}^{RC} may all be zero, resulting in an independence model. At another extreme, μ_{ij}^{RC} may be “saturated,” taking $(I - 1)(J - 1)$ degrees of freedom, yielding exact predictions ($F_{ij} = f_{ij}$ for all i and j).

Typically, the researcher is interested in fitting models between the two extreme cases by altering specifications for μ^{RC} . It is easy to show that all ODDS RATIOS in a two-way table are functions of the interaction parameters (μ^{RC}). Let θ_{ij} denote a local log-odds ratio for a 2×2 subtable formed from four adjacent cells obtained from two adjacent row categories and two adjacent column categories:

$$\theta_{ij} = \log\{[F_{(i+1)(j+1)}F_{ij}]/[F_{(i+1)j}F_{i(j+1)}]\},$$

$$i = 1, \dots, I-1, j = 1, \dots, J-1.$$

Let us assume that the row and column variables are ordinal on some scales x and y . The scales may be observed or latent. A linear-by-linear association model is as follows:

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + \beta x_i y_j, \quad (2)$$

where β is the parameter measuring the association between the two scales x and y representing, respectively, the row and column variables. If the two scales x and y are directly observed or imputed from external sources, estimation of equation (2) is straightforward via MAXIMUM LIKELIHOOD ESTIMATION for log-linear models.

ASSOCIATION MODELS FOR A TWO-WAY TABLE

If we do not have extra information about the two scales x and y , we can either impose assumptions about the scales or estimate the scales internally. Different approaches give rise to different association models. Below, I review the most important ones.

Uniform Association

If the categories of the variables are correctly ordered, the researcher may make a simplifying assumption that the ordering positions form the scales (i.e., $x_i = i$, $y_j = j$). Let the practice be called *integer scoring*. The integer-scoring simplification results in the following uniform association model:

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + \beta_{ij}. \quad (3)$$

The researcher can estimate the model with actual data to see whether this assumption holds true.

Row Effect and Column Effect Models

Although the uniform association model is based on integer scoring for both the row and column variables,

the researcher may wish to invoke integer scoring only for the row *or* the column variable. When integer scoring is used only for the column variable, the resulting model is called the *row effect model*. Conversely, when integer scoring is used only for the row variable, the resulting model is called the *column effect model*. Taking the row effect model as an example, we can derive the following model from equation (2):

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + j\phi_i. \quad (4)$$

This model is called the row effect model because the latent scores of the row variable ($\phi_i = \beta x_i$) are revealed by estimation after we apply integer scoring for the column variable. That is, ϕ_i is the “row effect” on the association between the row variable and the column variable. Note that the terms *row effect* and *column effect* here have different meanings than μ_i^R and μ_j^C , which are fitted to saturate the marginal distributions of the row and column variables.

Goodman's RC Model

The researcher can take a step further and treat both the row and column scores as unknown. Two of Goodman's (1979) association models are designed to estimate such latent scores. Goodman's Association Model I simplifies equation (1) to the following:

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + j\phi_i + i\varphi_j, \quad (5)$$

where ϕ_i and φ_j are, respectively, unknown row and column scores as in the row effect and column effect models. However, it is necessary to add three normalization constraints to uniquely identify the $(I + J)$ unknown parameters of ϕ_i and φ_j .

Goodman's Association Model I requires that both the row and column variables be correctly ordered a priori because integer scoring is used for both, as shown in equation (5). This requirement means that the model is not invariant to positional changes in the categories of the row and column variables. If the researcher has no knowledge that the categories are correctly ordered or in fact needs to determine the correct ordering of the categories, Model I is not appropriate. For this reason, Goodman's Association Model II has received the most attention. It is of the following form:

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + \beta\phi_i\varphi_j, \quad (6)$$

where β is the association parameter, and ϕ_i and φ_j are unknown scores to be estimated. Also, ϕ_i and φ_j are

Table 1 Comparison of Association Models

Model	μ^{RC}	DF_m	θ_{ij}
Uniform association	β_{ij}	1	β
Row effect	$j\phi_i$	$(I - 1)$	$\phi_{i+1} - \phi_i$
Column effect	$i\phi_j$	$(J - 1)$	$\phi_{j+1} - \phi_j$
Association Model I	$j\phi_i + i\phi_j$	$I + J - 3$	$(\phi_{i+1} - \phi_i) + (\phi_{j+1} - \phi_j)$
Association Model II (RC)	$\beta\phi_i\phi_j$	$I + J - 3$	$(\phi_{i+1} - \phi_i)(\phi_{j+1} - \phi_j)$

subject to four normalization constraints because each requires the normalization of both location and scale.

As equation (6) shows, the interaction component (μ^{RC}) of Goodman's Association Model II is in the form of multiplication of unknown parameters—log-bilinear specification. The model is also known as the *log-multiplicative model*, or simply the RC model. The RC model is very attractive because it allows the researcher to estimate unknown parameters even when the categories of the row and the column variables may not be correctly ordered. All that needs to be assumed is the existence of the ordinal scales. The model can reveal the orders through estimation.

Table 1 presents a summary comparison of the aforementioned association models. The second column displays the model specification for the interaction parameters (μ^{RC}). The number of degrees of freedom for each μ^{RC} specification is given in the third column (DF_m). If there are no other model parameters to be estimated, the degrees of freedom for a model are equal to $(I - 1)(J - 1) - DF_m$. The formula for calculating the local log-odds ratio is shown in the last column.

Goodman's Association Model II (RC model) can be easily extended to have multiple latent dimensions so that μ^{RC} of equation (1) is specified as

$$\mu_{ij}^{RC} = \sum \beta_m \phi_{im} \phi_{jm}, \quad (7)$$

where the summation sign is with respect to all possible m dimensions, and the parameters are subject to necessary normalization constraints. Such models are called $RC(M)$ models. See Goodman (1986) for details.

ASSOCIATION MODELS FOR THREE-WAY AND HIGHER-WAY TABLES

Below, I mainly discuss the case of a three-way table. Generalizations to a higher-way table can be easily made. Let R denote row, C denote column, and L denote layer, with their categories indexed

respectively by i ($i = 1, \dots, I$), j ($j = 1, \dots, J$), and k ($k = 1, \dots, K$). In a common research setup, the researcher is interested in understanding how the two-way association between R and C depends on levels of L . For example, in a trend analysis, L may represent different years or cohorts. In a comparative study, L may represent different nations or groups. Thus, research attention typically focuses on the association pattern between R and C and its variation across layers.

Let F_{ijk} denote the expected frequency in the i th row, the j th column, and the k th layer. The saturated log-linear model can be written as follows:

$$\log(F_{ijk}) = \mu + \mu_i^R + \mu_j^C + \mu_k^L + \mu_{ij}^{RC} + \mu_{ik}^{RL} + \mu_{jk}^{CL} + \mu_{ijk}^{RCL}. \quad (8)$$

In a typical research setting, interest centers on the variation of the RC association across layers. Thus, the baseline (for the null hypothesis) is the following conditional independence model:

$$\log(F_{ijk}) = \mu + \mu_i^R + \mu_j^C + \mu_k^L + \mu_{ik}^{RL} + \mu_{jk}^{CL}. \quad (9)$$

That is, the researcher needs to specify and estimate μ^{RC} and μ^{RCL} to understand the layer-specific RC association.

There are two broad approaches to extending association models to three-way or higher-way tables. The first is to specify an association model for the typical association pattern between R and C and then estimate parameters that are specific to layers or test whether they are invariant across layers (Clogg, 1982a). The general case of the approach is to specify μ^{RC} and μ^{RCL} in terms of the RC model so as to change equation (8) to the following:

$$\log(F_{ijk}) = \mu + \mu_i^R + \mu_j^C + \mu_k^L + \mu_{ik}^{RL} + \mu_{jk}^{CL} + \beta_k \phi_{ik} \phi_{jk}. \quad (10)$$

That is, the β , ϕ , and φ parameters can be layer specific or layer invariant, subject to model specification and statistical tests. The researcher may also wish to test special cases (i.e., the uniform association, column effect, and row effect models) where ϕ and/or φ parameters are inserted as integer scores rather than estimated.

The second approach, called the *log-multiplicative layer-effect model*, or the “unidiff model,” is to allow a flexible specification for the typical association pattern between R and C and then to constrain its cross-layer variation to be log-multiplicative (Xie, 1992). That is, we give a flexible specification for μ^{RC} but constrain μ^{RCL} so that equation (8) becomes the following:

$$\log(F_{ijk}) = \mu + \mu_i^R + \mu_j^C + \mu_k^L + \mu_{ik}^{RL} + \mu_{jk}^{CL} + \phi_k \psi_{ij}. \quad (11)$$

With the second approach, the RC association is not constrained to follow a particular model and indeed can be saturated with $(I - 1)(J - 1)$ dummy variables. In a special case in which the typical association pattern between R and C is the RC model, the two approaches coincide, resulting in the three-way RCL log-multiplicative model. Powers and Xie (2000, pp. 140–145) provide a more detailed discussion of the variations and the practical implications of the second approach. It should be noted that the two approaches are both special cases of a general framework proposed by Goodman (1986) and extended in Goodman and Hout (1998).

APPLICATIONS

Association models have been used widely in sociological research. Below, I give three concrete examples. The first example is one of scaling. See Clogg (1982b) for a detailed illustration of this example. Clogg aimed to scale an ordinal variable that measures attitude on abortion. The variable was constructed from a Guttman scale, and the cases that did not conform to the scale response patterns were grouped into a separate category, “error responses.” As is usually the case, scaling required an “instrument.” In this case, Clogg used a measure of attitude on premarital sex that was collected in the same survey. The underlying assumption was that the scale of the attitude on abortion could be revealed from its association with the attitude on premarital sex. Clogg used the log-multiplicative model to estimate the scores associated with the different categories of the two variables. Note that the log-multiplicative RC

model assumes that the categories are ordinal but not necessarily correctly ordered. So, estimation reveals the scale as well as the ordering. Through estimation, Clogg showed that the distances between the adjacent categories were unequal and that those who gave “error responses” were in the middle in terms of their attitudes on abortion.

The second example is the application of the log-multiplicative layer-effect model to the cross-national study of intergenerational mobility (Xie, 1992). The basic idea is to force cross-national differences to be summarized by layer-specific parameters [i.e., ϕ_k of equation (11)] while allowing and testing different parameterizations of the two-way association between father’s occupation and son’s occupation (i.e., ψ_{ij}). The ϕ_k parameters are then taken to represent the social openness or closure of different societies.

The third example, which involves the study of human fertility, is nonconventional in the sense that the basic setup is not log-linear but log-rate. The data structure consists of a table of frequencies (births) cross-classified by age and country and a corresponding table of associated exposures (women-years). The ratio between the two yields the country-specific and age-specific fertility rates. The objective of statistical modeling is to parsimoniously characterize the age patterns of fertility in terms of fertility level and fertility control for each country. In conventional demography, this is handled using Coale and Trussell’s Mm method. Xie and Pimentel (1992) show that this method is equivalent to the log-multiplicative layer-effect model, with births as the dependent variable and exposure as an “offset.” Thus, the M and m parameters of Coale and Trussell’s method can be estimated statistically along with other unknown parameters in the model.

ESTIMATION

Estimation is straightforward with the uniform, row effect, column effect, and Goodman’s Association Model I models. The user can use any of the computer programs that estimate a log-linear model. What is complicated is when the RC interaction takes the form of the product of unknown parameters—the log-multiplicative or log-bilinear specification. In this case, a reiterative estimation procedure is required. The basic idea is to alternately treat one set of unknown parameters as known while estimating the other and to continue the iteration process until both are stabilized. Special computer programs, such as ASSOC and

CDAS, have been written to estimate many of the association models. User-written subroutines in GLIM and STATA are available from individual researchers. For any serious user of association models, I also recommend Lem, a program that can estimate different forms of the log-multiplicative model while retaining flexibility. See my Web site www.yuxie.com for updated information on computer subroutines and special programs.

—Yu Xie

REFERENCES

- Clogg, C. C. (1982a). Some models for the analysis of association in multiway cross-classifications having ordered categories. *Journal of the American Statistical Association*, 77, 803–815.
- Clogg, C. C. (1982b). Using association models in sociological research: Some examples. *American Journal of Sociology*, 88, 114–134.
- Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74, 537–552.
- Goodman, L. A. (1984). *The analysis of cross-classified data having ordered categories*. Cambridge, MA: Harvard University Press.
- Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review*, 54, 243–309.
- Goodman, L. A., & Hout, M. (1998). Understanding the Goodman-Hout approach to the analysis of differences in association and some related comments. In Adrian E. Raftery (Ed.), *Sociological methodology* (pp. 249–261). Washington, DC: American Sociological Association.
- Powers, D. A., & Xie, Y. (2000). *Statistical methods for categorical data analysis*. New York: Academic Press.
- Xie, Y. (1992). The log-multiplicative layer effect model for comparing mobility tables. *American Sociological Review*, 57, 380–395.
- Xie, Y., & Pimentel, E. E. (1992). Age patterns of marital fertility: Revising the Coale-Trussell method. *Journal of the American Statistical Association*, 87, 977–984.

ASSUMPTIONS

Assumptions are ubiquitous in social science. In theoretical work, assumptions are the starting axioms and postulates that yield testable implications spanning broad domains. In empirical work, statistical procedures typically embed a variety of assumptions, for example, concerning measurement

properties of the variables and the distributional form and operation of unobservables (such as HOMOSKEDASTICITY or NORMAL DISTRIBUTION of the ERROR). Assumptions in empirical work are discussed in the entries for particular procedures (e.g., ORDINARY LEAST SQUARES); here we focus on assumptions in theories.

The purpose of a scientific THEORY is to yield testable implications concerning the relationships between observable phenomena. The heart of the theory is its set of assumptions. The assumptions embody what Popper (1963) calls “guesses” about nature—guesses to be tested, following Newton’s vision, by testing their logical implications. An essential feature of the assumption set is internal logical consistency. In addition, three desirable properties of a theory are as follows: (a) that its assumption set be as short as possible, (b) that its observable implications be as many and varied as possible, and (c) that its observable implications include phenomena or relationships not yet observed, that is, novel predictions.

Thus, a theory has a two-part structure: a small part containing the assumptions and a large and ever-growing part containing the implications. Figure 1 provides a visualization of the structure of a theory.

A theory can satisfy all three properties above and yet be false. That is, one can invent an imaginary world, set up a parsimonious set of postulates about its operation, deduce a wide variety of empirical consequences, and yet learn through empirical test that no known world operates in conformity with the implications derived from the postulated properties of the imaginary world. That is why empirical analysis is necessary, or, put differently, why theoretical analysis alone does not suffice for the accumulation of reliable knowledge.

A note on terminology: *Assumption* is a general term, used, as noted earlier, in both theoretical and empirical work. Sharper terms sometimes used in theoretical work include *axiom*, which carries the connotation of *self-evident*, and *postulate*, which does not, and is therefore more faithful to an enterprise marked by guesses and bound for discovery. Other terms include the serviceable *premise*, the colorful *starting principle*, and the dangerous *hypothesis*, which is used not only as a postulate (as in the HYPOTHETICO-DEDUCTIVE METHOD invented by Newton) but also as an observable proposition to be tested.

Where do assumptions come from? Typically, the frameworks for analyzing topical domains include a variety of relations and FUNCTIONS, some of which may prove to be fruitful assumptions. In general,

PREDICTIONS

Figure 1 Structure of a Theory

mathematically expressed functions make promising assumptions, as (a) their very statement signals generality and breadth, and (b) they are amenable to manipulation via mathematical tools, which make it easy to derive a wealth of implications and, unlike verbal covering-law procedures, make it possible to derive novel predictions.

The tools for derivation of predictions include the full panoply of mathematical tools. Besides functions, these include DISTRIBUTIONS, MATRICES, inequalities, and partial differential equations. For example, the quantities related via functions can be characterized by their distributional properties, and the distributional form shapes the character of the predictions. The results can be surprising. Assumptions that look innocent turn out to touch vast domains in unexpected ways. To illustrate, the justice evaluation function, when used as an assumption, yields predictions for parental gift giving, differential mourning of mothers and fathers, differential risk of posttraumatic stress among combat veterans, differential social contributions of monastic and mendicant religious institutions, and inequality effects on migration (Jasso, 2001).

To assess the usefulness of assumptions, it is convenient to maintain a spreadsheet showing, for a particular theory, which of the assumptions in its assumption set are used in the derivation of each prediction. This layout tells at a glance which assumptions are doing most of the work and how many assumptions each prediction requires. Figure 2 provides a visualization of this type of layout, called a Merton Chart of Theoretical Derivation.

A theory and its assumptions may be more, or less, fundamental (BASIC RESEARCH). The Holy Grail is discovery of ultimate forces governing the socio-behavioral world. The theories and assumptions we work with currently, even the most fruitful and the ones with the longest reach, are like signposts along the road. But what signposts. Simultaneously, they assist our preparation and sharpening of tools for the deeper theories to come and illuminate patches of the road.

FURTHER READING

Cogent background in the philosophy of science is provided by Toulmin (1978). Elaboration of procedures for constructing sociobehavioral theories, including choice of assumptions and their mathematical manipulation, is provided in Jasso (1988, 2001, 2002). More fundamentally and more important, it is

instructive, not to mention inspiring, to watch great theorists in action, choosing assumptions, refining them, deriving implications from them. A good starting point is Newton's (1686/1952) *Principia*, in which he worked out the method subsequently called the hypothetico-deductive method. Of course, the reader's taste will dictate the best starting point. Other possibilities include Mendel's (1866) paper, which laid the foundations for genetics, and the little Feynman and Weinberg (1987) volume, which also includes a foreword recalling Dirac's prediction of antimatter.

—Guillermina Jasso

REFERENCES

- Feynman, R. P., & Weinberg, S. (1987). *Elementary particles and the laws of physics: The 1986 Dirac Memorial Lectures*. Cambridge, UK: Cambridge University Press.
- Jasso, G. (1988). Principles of theoretical analysis. *Sociological Theory*, 6, 1–20.
- Jasso, G. (2001). Formal theory. In J. H. Turner (Ed.), *Handbook of sociological theory* (pp. 37–68). New York: Kluwer Academic/Plenum.
- Jasso, G. (2002). Seven secrets for doing theory. In J. Berger & M. Zelditch (Eds.), *New directions in contemporary sociological theory* (pp. 317–342). Boulder, CO: Rowan & Littlefield.
- Mendel, G. (1866). Versuche über Pflanzen-Hybriden [Experiments in plant hybridization]. In *Verhandlungen des naturforschenden Vereins* [Proceedings of the Natural History Society]. Available in both the original German and the English translation at www.mendelweb.org
- Newton, I. (1952). *Mathematical principles of natural philosophy*. Chicago: Britannica. (Original work published 1686.)
- Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. New York: Basic Books.
- Toulmin, S. E. (1978). Science, philosophy of. In *Encyclopaedia Britannica* (15th ed., Vol. 16, pp. 375–393). Chicago: Britannica.

ASYMMETRIC MEASURES

ASSOCIATION coefficients measure the strength of the relation between two variables X and Y and, for ordinal and quantitative (interval and ratio-scaled) variables, the direction of the association. Asymmetric (or directional) association coefficients also assume that one of the two variables (e.g., X) can be identified as the independent variable and the other variable (e.g., Y) as the dependent variable.

As an example, in a survey of university graduates, the variable X might be gender (nominal, dichotomous), ethnic group (nominal), the socioeconomic status of parents (ordinal), or the income of parents (quantitative). The variable Y might be the field of study or party affiliation (nominal), a grade (ordinal), the income of respondents (quantitative), or membership in a union (nominal, dichotomous, with a “yes” or “no” response category).

Available Coefficients

Most statistical software packages, such as SPSS, offer asymmetric coefficients only for special combinations of measurement levels. These coefficients are described first.

Measures for Nominal-Nominal Tables

Measures for nominal variables are lambda (λ), Goodman and Kruskal's tau (τ), and the uncertainty coefficient U . All measures are based on the definition of PROPORTIONAL REDUCTION OF ERROR (PRE):

$$\text{PRE} = \frac{E_0 - E_1}{E_0}, \quad (1)$$

where E_0 is the error in the dependent variable Y without using the independent variable X to predict or explain Y , and E_1 is the error in Y if X is used to predict or explain Y .

Both τ and U use definitions of variance to compute E_0 and E_1 . Similar to the well-known eta (η) from the ANALYSIS OF VARIANCE (see below), E_0 is the total variance of Y , and E_1 is the residual variance (variance within the categories of X). For τ , the so-called Gini concentration is used to measure the variation (Agresti, 1990, pp. 24–25), and U applies the concept of entropy. For λ , a different definition is used: It analyzes to what extent the dependent variable can be predicted both without knowing X and with knowledge of X .

All three coefficients vary between 0 (no reduction in error) and 1 (perfect reduction in error). Even if a clear association exists, λ may be zero or near zero. This is the case if one category of the dependent variable is dominant, so that the same category of Y has the highest frequencies and becomes the best predictor within each condition X .

Measure for Ordinal-Ordinal Tables

The most prominent measure for this pattern is Somers' d . Somers' d is defined as the least squares regression slope ($d_{y/x} = s_{xy}/s_x^2$) between the dependent variable Y and the independent variable X , if both variables are treated as ordinal (Agresti, 1984, pp. 161–163), and Daniels' formula for generalized correlation coefficient (see ASSOCIATION) is used to compute the variance s_x^2 of X and the covariance s_{xy} between X and Y . Somers' d is not a PRE coefficient. The corresponding PRE coefficient is the symmetric Kendall's τ_b^2 , which can be interpreted as explained variance and is equal to the geometric mean if Somers' d is computed for X and Y as dependent variables ($= d_{y/x}$ and $d_{x/y}$). If both variables are dichotomous, Somers' d is equal to the difference of proportions (Agresti, 1984, p. 161).

Measures for Nominal-Interval (or Ratio) Tables

The well-known explained variance η^2 or, respectively, η (eta) from ANALYSIS OF VARIANCE (ANOVA) can be applied to situations in which the independent variable is nominal and the dependent variable is quantitative. η^2 is a PRE coefficient.

Formulas

Formulas for computing the described measures are given in Agresti (1990) and SPSS (2002), for example.

Measures for Other Combinations

Up until now, only pairings for nominal-nominal tables (= Constellations 1, 2, 5, and 6 in Table 1), ordinal-ordinal tables (= Constellation 11), and nominal-quantitative tables (= Constellations 4 and 8) have been analyzed.

Strategies to obtain measures for the other pairings, as well as those already analyzed, include the following:

- *Use of symmetric measures.* Association measures can be derived using Daniels's generalized correlation coefficient for all constellations (1, 3, 4, 9, 11 to 13, 15, 16), except those having one or two nominal-polytomous variables (Constellations 2, 5 to 8, 10, 14).

Table 1 Possible Pairings of Measurement Levels for Two Variables

Independent Variable <i>X</i>	Dependent Variable <i>Y</i>			
	Nominal-Dichotomous	Nominal-Polytomous	Ordinal	Quantitative (Interval or Ratio)
Nominal-dichotomous	1	2	3	4
Nominal-polytomous	5	6	7	8
Ordinal	9	10	11	12
Quantitative (interval or ratio)	13	14	15	16

• *Reduction of measurement level.* The measurement level of the variable with the higher measurement level is reduced. If, for example, *X* is nominal (polytomous) and *Y* is ordinal (Constellation 7), it is recommended to treat *Y* as nominal scaled (Constellation 6) and apply λ , τ , or U as an appropriate measure. This procedure results in a loss of information and can hide an association.

• *Upgrading of measurement level.* Ordinal variables are treated quantitatively. From a strict theoretical point of view, this strategy is incompatible. However, textbooks (see below) recommend this strategy, which is also frequently applied in survey research. In many surveys, I very often have found a nearly perfect linear relation between ordinal and quantitative measures. But convince yourself.

• *Use of logistic regression.* Appropriate LOGISTIC REGRESSION models have been developed for all constellations (Pairings 1 to 3, 5 to 7, 9 to 11, 13 to 15), except for those in which the dependent variable is quantitative (Pairings 4, 8, 12, and 16) (Agresti, 1984, pp. 104–131). Ordinal independent variables are treated as quantitative variables (Agresti, 1984, p. 114) or as nominal scaled.

• *Use of linear regression.* LINEAR REGRESSION can be used for Pairing 16 as well as for Pairings 4 and 8. For Constellation 8, dummies must be computed.

• *Use of special measures.* Special approaches have been developed: ridit analysis (Agresti, 1984, pp. 166–168) and the dominance statistic Δ (Agresti, 1984, pp. 166–168) can be applied to Constellations 3 and 7, as well as to 11 and 15, if the independent variable is treated as nominal scaled.

Dichotomous Variables. These take an exceptional position. They can be treated either as ordinal or quantitative. Therefore, all strategies for Constellation

11 can also be applied to Constellations 1, 3, and 9, and all strategies for Constellation 16 can be applied to Constellations 1, 4, and 13 and so on.

Generalizations. The great advantage to using logistic or linear regression is that generalizations exist when there is more than one independent variable. Hence, it is recommended to use these advanced techniques. However, logistic regression models do not solve the problem of interpreting the relation's strength. As with all association coefficients, general threshold values are not possible (see ASSOCIATION).

Significance of Coefficients. Test statistics are available for all measures and coefficients discussed. Computational formulas are given in SPSS (2002).

—Johann Bacher

REFERENCES

- Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: John Wiley.
- Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley.
- SPSS. (2002, August). Crosstabs. In *SPSS 11.0 statistical algorithms* [Online]. Available: <http://www.spss.com/tech/stat/algorithms/11.0/crosstabs.pdf>

ASYMPTOTIC PROPERTIES

To make a STATISTICAL INFERENCE, we need to know the SAMPLING DISTRIBUTION of our ESTIMATOR. In general, the exact or finite sample distribution will not be known. The asymptotic properties provide us with the approximate distribution of our estimator by looking at its behavior as a data set gets arbitrarily large (i.e., infinite).

Asymptotic results are based on the sample size n . Let x_n be a RANDOM VARIABLE indexed by the size of the SAMPLE. To define consistency, we will need to define the notation of convergence in PROBABILITY. We will say that x_n converges in probability to a constant c if $\lim_{n \rightarrow \infty} \text{Prob}(|x_n - c| > \varepsilon) = 0$ for any positive ε . Convergence in probability implies that the values that the variable may take that are not close to c become increasingly unlikely as the sample size n increases. Convergence in probability is usually denoted by $\text{plim } x_n = c$. We can then say that an ESTIMATOR $\hat{\theta}_n$ of a parameter θ is a consistent estimator if and only if $\text{plim } \hat{\theta}_n = \theta$. Roughly speaking, this means that as the sample size gets large, the estimator's sampling DISTRIBUTION piles up on the true value.

Although consistency is an important characteristic of an estimator, to conduct statistical inference, we will need to know its sampling distribution. To derive an estimator's sampling distribution, we will need to define the notation of a limiting distribution. Suppose that $F(x)$ is the cumulative distribution function (cdf) of some random variable x . Furthermore, let x_n be a sequence of random variables indexed by sample size with cdf $F_n(x)$. We say that x_n converges in distribution to x if $\lim_{n \rightarrow \infty} |F_n(x) - F(x)| = 0$.

We will denote convergence in distribution by $x_n \xrightarrow{d} x$. In words, this says that the distribution of x_n gets close to the distribution of x as the sample size gets large. This convergence in distribution, along with the CENTRAL LIMIT THEOREM, will allow us to construct the asymptotic distribution of an estimator.

An asymptotic distribution is a distribution that is used to approximate the true finite sample distribution of a random variable. Suppose that $\hat{\theta}_n$ is a consistent estimate of a parameter θ . Then, using the central limit theorem, the asymptotic distribution of $\hat{\theta}_n$ is

$$\sqrt{n}(\hat{\theta}_n - \theta)^d \rightarrow N[0, V], \quad (1)$$

where $N[]$ denotes the cdf of the NORMAL DISTRIBUTION, and V is the asymptotic covariance matrix. This implies that $\hat{\theta}_n$ is distributed as $N[\theta, \frac{1}{n}V]$. This is a very striking result. It says that, regardless of the true underlying finite sample distribution, we can approximate the distribution of the estimator by the Normal distribution as the sample size gets large. We can thus use the Normal distribution to conduct HYPOTHESIS TESTING as if we knew the true sampling distribution.

However, it must be noted that the asymptotic distribution is only an approximation, although one that will improve with sample size; thus, the size, power, and other properties of the hypothesis test are only approximate. Unfortunately, we cannot, in general, say how large a sample needs to be to have confidence in using the asymptotic distribution for statistical inference. One must rely on previous experience, BOOTSTRAPPING simulations, or MONTE CARLO experiments to provide an indication of how useful the asymptotic approximations are in any given situation.

—Jonathan N. Katz

REFERENCE

White, H. (1984). *Asymptotic theory for econometricians*. Orlando, FL: Academic Press.

ATLAS.ti

ATLAS.ti is a software program designed for computer-assisted qualitative data analysis (see CAQ-DAS). As such, it allows qualitative data to be coded and retrieved. In addition, it supports the analysis of visual images as well as text. It also supports GROUNDED THEORY practices, such as the generation of MEMOS.

—Alan Bryman

ATTENUATION

Attenuation refers to the CORRELATION between two different measures being reduced due to measurement error. Of course, no test has scores that are perfectly reliable. Therefore, having unreliable scores results in poor predictability of the criterion. To obtain an estimate of the correlation, given predictor and criterion scores, with perfect RELIABILITY, Spearman (1904) proposed the following formula known as the correction for attenuation:

$$\bar{r}_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}}. \quad (1)$$

This formula has also been referred to as the double correction because both the predictor (x) and criterion

(y) scores are being corrected. Suppose that we are examining the correlation between depression and job satisfaction. The correlation between the two measures (r_{xy}) is .20. If the reliability (e.g., CRONBACH'S ALPHA) of the scores from the depression inventory (r_{xx}) is .80 and the reliability of the scores from the job satisfaction scale (r_{yy}) is .70, then the correlation corrected for attenuation would be equal to the following:

$$\bar{r}_{xy} = \frac{.20}{\sqrt{.80 \times .70}} = .26. \quad (2)$$

In this case, there is little correction. However, suppose the reliabilities of the scores for depression and job satisfaction are .20 and .30, respectively. The correlation corrected for attenuation would now be equal to .81. The smaller the reliabilities or sample sizes ($N < 300$), the greater the correlation corrected for attenuation (Nunnally, 1978). It is even possible for the correlation corrected for attenuation to be greater than 1.00 (Nunnally, 1978).

There are cases, however, when one might want to correct for unreliability for either the predictor or the criterion scores. These would incorporate the single correction. For example, suppose the correlation between scores on a mechanical skills test (x) and job performance ratings (y) is equal to .40, and the reliability of the job performance ratings is equal to .60. If one wanted to correct for only the unreliability of the criterion, the following equation would be used:

$$\bar{r}_{xy} = \frac{r_{xy}}{\sqrt{r_{yy}}}. \quad (3)$$

In this case, the correlation corrected for attenuation would be equal to .51. It is also feasible, albeit less likely for researchers, to correct for the unreliability of the test scores only. This correction would occur when either the criterion reliability is measured perfectly or if the criterion reliability is unknown (Muchinsky, 1996). Hence, the following formula would be applied:

$$\bar{r}_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}}}. \quad (4)$$

Muchinsky (1996) summarized the major tenets of the correction for attenuation. First, the correction for attenuation does not increase the predictability of the test scores (Nunnally, 1978). Furthermore, the corrected correlations should neither be tested for STATISTICAL SIGNIFICANCE (Magnusson, 1967), nor should

corrected and uncorrected VALIDITY coefficients be compared to each other (American Educational Research Association, 1985). Finally, the correction for attenuation should not be used for averaging different types of validity coefficients in META-ANALYSIS that use different estimates of reliabilities (e.g., internal consistencies, test-retest) (Muchinsky, 1996).

—N. Clayton Silver

REFERENCES

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.
- Magnusson, D. (1967). *Test theory*. Reading, MA: Addison-Wesley.
- Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, 56, 63–75.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.

ATTITUDE MEASUREMENT

An attitude is an evaluation of an object (person, place, or thing) along a dimension ranging from favorable to unfavorable (Weiss, 2002). As such, an attitude has both affective (emotional) and cognitive aspects. Because they are internal states of people, attitudes can best be assessed via self-reports with quantitative psychological scales that can have several different formats (Judd, Smith, & Kidder, 1991).

THURSTONE SCALE

A THURSTONE SCALE (Thurstone & Chave, 1929) consists of a list of items varying in favorability toward the attitude object. The underlying idea is that an individual will agree with only those items that are close in favorability to where his or her attitude lies. For example, Spector, Van Katwyk, Brannick, and Chen (1997) created a scale of job attitudes that had items that ranged from extreme negative (e.g., *I hate my job*) to extreme positive (e.g., *My*

Table 1 Example of a Semantic Differential

<i>The Supreme Court of the United States</i>								
Good	_____	_____	_____	_____	_____	_____	_____	Bad
Fair	_____	_____	_____	_____	_____	_____	_____	Unfair
Weak	_____	_____	_____	_____	_____	_____	_____	Strong
Passive	_____	_____	_____	_____	_____	_____	_____	Active
Happy	_____	_____	_____	_____	_____	_____	_____	Unhappy
Unimportant	_____	_____	_____	_____	_____	_____	_____	Important

job is the most enjoyable part of my life), with more moderate positive (e.g., *I enjoy my job*) and negative items (e.g., *I don't really like my job very much*) in between.

A Thurstone scale is developed by first writing a number of items that are given to a panel of judges to sort into categories that vary from extremely unfavorable to extremely favorable. The categories are numbered, and the mean score across the judges is computed. To use the scale, respondents are asked to read the items presented in a random order and indicate for each whether they agree or not. The attitude score is the mean scale value of the items the individual endorsed.

LIKERT (SUMMATED RATING) SCALE

The LIKERT or SUMMATED RATING SCALE (Likert, 1932) consists of a series of statements that are either favorable or unfavorable about the attitude object. Respondents are given multiple response choices that vary along a continuum, usually agree to disagree, and they are asked to indicate their extent of agreement with each one. The scale is designed on the assumption that the favorable and unfavorable items are mirror images of one another, so that an individual who tends to agree with one set of items will disagree with the other. This is in contrast to the assumption of the Thurstone scale that people agree only with items that are close in favorability or unfavorability to their own attitude position. Spector et al. (1997) demonstrated that the Likert scale assumption is not always correct, and this can produce artifactual results when such items are factor analyzed.

The Likert scale has become the most popular for the assessment of attitudes, undoubtedly for two major reasons. First, it is relatively easy to develop (see LIKERT SCALE for the procedure), requiring no

initial scaling study to get judges' ratings of the items. Second, this approach has been shown to yield reliable scales that are useful for assessing attitudes.

SEMANTIC DIFFERENTIAL

The SEMANTIC DIFFERENTIAL SCALE was developed by Osgood, Suci, and Tannenbaum (1957) to assess different aspects of attitude toward an object. The stem for a semantic differential is the name or description of the attitude object (e.g., Supreme Court of the United States). A series of bipolar adjectives is provided with a series of response choices in between. The respondent is asked to check the choice that best reflects his or her feeling, where the nearer the choice is to the adjective, the stronger the person endorses it (see Table 1).

The semantic differential items tend to group into three underlying dimensions of attitude: evaluation, potency, and activity. Evaluation indicates the extent to which the person has a favorable or unfavorable attitude, that is, whether he or she likes or dislikes the object. Potency is the strength of the object, which indicates the extent to which a person sees the attitude object as being powerful or not, regardless of favorability. Activity is the extent to which the attitude object is seen as being active or passive.

Semantic differentials tend to be used to assess attitudes about several objects at the same time rather than to develop an attitude scale for one object. It is possible to include only evaluation bipolar adjectives to yield a single score per attitude object as the sum of the scores across items. This can be an easy and efficient procedure for respondents, as there is less reading with this sort of scale than the Thurstone or Likert scale, and the same items can be used for each attitude object, making comparisons easier to make.

GUTTMAN SCALE

A GUTTMAN SCALE (Guttman, 1944) consists of a hierarchy of items that vary from unfavorable to favorable. Respondents are asked to indicate for each item whether he or she agrees with it. It is assumed that individuals with favorable attitudes will agree with all favorable items that are at their level or lower. Thus, a slightly favorable person will agree with a slightly favorable item but not an extremely favorable item, whereas an extremely favorable person will agree with both slightly and extremely favorable items. The Guttman scale has not been used as frequently as the other scales, probably for two reasons. First, the scale can be more difficult to develop than the more popular Likert and semantic differential scales. Second, the expected pattern of responses does not always occur because often those with extreme attitudes might disagree with slightly favorable items, or the items might reflect more than a single attitude dimension.

—Paul E. Spector

REFERENCES

- Guttman, L. (1944). A basis for scaling quantitative data. *American Sociological Review*, 9, 139–150.
- Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). Fort Worth, TX: Harcourt Brace Jovanovich.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 55.
- Osgood, C. E., Suci, C. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.
- Spector, P. E., Van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management*, 23, 659–677.
- Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.
- Weiss, H. M. (2002). Deconstructing job satisfaction: Separating evaluations, beliefs and affective experiences. *Human Resources Management Review*, 12, 173–194.

ATTRIBUTE

One definition of *attribute* is a characteristic that a respondent, subject, or case does or does not have. A binary attribute variable has two values: the presence of the attribute and its absence. For example, adults in a survey may have the attribute of

being married (score = 1) or not (score = 0). Another definition, from psychological or marketing research, refers to attributes as scalable items. In decision research, for example, attributes may refer to goals or criteria relevant to the decision maker. To illustrate, in deciding on a new car, the buyer may be faced with three alternatives and have to weigh attributes such as fuel efficiency, cost, and safety.

—Michael S. Lewis-Beck

See also BINARY, DICHOTOMOUS VARIABLES

ATTRITION

Attrition is a term used to describe the process by which a SAMPLE reduces in size over the course of a survey data collection process due to NONRESPONSE and/or due to units ceasing to be eligible. The term is also used to refer to the numerical outcome of the process.

In PANEL surveys, the number of sample units having responded at every wave decreases over waves due to the cumulative effects of nonresponse and changes in eligibility status. Such attrition may introduce unquantifiable bias into survey estimates due to the nonrandom nature of the dropout. For this reason, considerable efforts are often made to minimize the extent of sample attrition. The extent of attrition is affected also by factors such as the number of waves of data collection, the intervals between waves, the burden of the data collection exercise, and so on. WEIGHTING and other adjustment methods may be used at the analysis stage to limit the detrimental impacts of attrition.

Although perhaps most pronounced in the context of panel surveys, attrition can affect any survey with multiple stages in the data collection process. Many CROSS-SECTIONAL surveys have multiple data collection instruments—for example, a personal interview followed by a SELF-COMPLETION (“drop-off”) QUESTIONNAIRE or an interview followed by medical examinations. In such cases, there will be attrition between the stages: Not all persons interviewed will also complete and return the self-completion questionnaire, for example. Even with a single data collection instrument, there may be multiple stages in the survey process. For example, in a survey of schoolchildren,

it may be necessary first to gain the cooperation of a sample of schools to provide lists of pupils from which a sample can be selected. It may then be necessary to gain parental consent before approaching the sample children. Thus, there are at least three potential stages of sample dropout: nonresponse by schools, parents, and pupils. This, too, can be thought of as a process of sample attrition.

In attempting to tackle attrition, either by reducing it at the source or by adjusting at the analysis stage, different component subprocesses should be recognized and considered explicitly. There is an important conceptual difference between attrition due to nonresponse by eligible units and attrition due to previously eligible units having become ineligible. The latter need not have a detrimental effect on analysis if identified appropriately. Each of these two main components of attrition splits into distinct subcomponents. Nonresponse could be due to a failure to contact the sample unit ("non-contact"), inability of the sample unit to respond (for reasons of language, illness, etc.), or unwillingness to respond ("refusal"). Change of eligibility status could be due to death, geographical relocation, change in status, and so forth. The different components have different implications for reduction and adjustment. Many panel surveys devote considerable resources to "tracking" sample members over time to minimize the noncontact rate. The experience of taking part in a wave of the survey (time taken, interest, sensitivity, cognitive demands, etc.) is believed to be an important determinant of the refusal rate at subsequent waves; consequently, efforts are made to maximize the perceived benefits to respondents of cooperation relative to perceived drawbacks.

—Peter Lynn

REFERENCES

- Laurie, H., Smith, R., & Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, 15(2), 269–282.
- Lepkowski, J. M., & Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 259–272). New York: John Wiley.
- Murphy, M. (1990). Minimising attrition in longitudinal studies: Means or end? In D. Magnusson & L. R. Bergman (Eds.), *Data quality in longitudinal research* (pp. 148–156). Cambridge, UK: Cambridge University Press.

AUDITING

Auditing is a means of establishing the dependability of findings in qualitative research. Dependability is associated with TRUSTWORTHINESS CRITERIA. Auditing entails keeping complete records of all phases of a qualitative research project and submitting them to peers who would assess whether proper procedures were followed, including the robustness of theoretical inferences and the ethical appropriateness of research practices.

—Alan Bryman

AUTHENTICITY CRITERIA

Authenticity criteria are criteria appropriate for judging the quality of inquiry (research, evaluation, and policy analysis) conducted within the protocols of emergent, nonpositivist paradigms such as CRITICAL THEORY, participant inquiry, CONSTRUCTIVISM, and others. Positivist criteria of rigor are well established: INTERNAL VALIDITY, EXTERNAL VALIDITY, RELIABILITY, and OBJECTIVITY. An early effort to devise similar but nonpositivist ways of assessing inquiry quality resulted in so-called TRUSTWORTHINESS CRITERIA (Guba, 1981), constructed as "parallels" to the four standard criteria and termed, respectively, *credibility*, *transferability*, *dependability*, and *confirmability*. But their very parallelism to the standard criteria of rigor makes their applicability to alternative inquiry approaches suspect. Moreover, the standard criteria are primarily *methodological*, overlooking such issues as power, ethics, voice, ACCESS, representation, and others raised by POSTMODERNISM.

The authenticity criteria were developed rooted in the propositions of alternative inquiry paradigms (Guba & Lincoln, 1989). This set is incomplete and primitive but may serve reasonably well as guidelines.

Certain *initial conditions* are prerequisite to all five authenticity criteria. These include the following: Respondents are drawn from all at-risk groups, fully INFORMED CONSENT procedures are in place, caring and trusting relationships are nurtured, inquiry procedures are rendered transparent to all respondents and audiences, and respondent-inquirer collaboration is built into every step, with full agreement on the rules to

govern the inquiry and with information fully shared. The inquiry report is guaranteed to be available to all respondents and audiences. Finally, an appellate mechanism is established to be used in cases of conflict or disagreement.

Definitions of the criteria, together with recommended procedures to establish them, follow:

- *Fairness* is defined as the extent to which all competing constructions of reality, as well as their underlying value structures, have been accessed, exposed, deconstructed, and taken into account in shaping the inquiry product, that is, the emergent reconstruction. Certain *procedures* should be strictly followed. All prior constructions of respondents and inquirer are obtained, compared, and contrasted, with each enjoying similar privilege; prior and later constructions are compared and contrasted; respondents and inquirer negotiate data to be collected, methods to be employed, interpretations to be made, modes of reporting to be employed, recommendations to be made, and actions to be proposed (taken); introspective statements (testimony) about changes experienced by respondents and inquirer are collected; prolonged engagement and persistent observation are used; individual and group member checks/validations are employed; thick contextual description is provided; peer debriefers and auditors are used (implying the maintenance of an audit trail); and, finally, the degree of empowerment felt by respondents is assessed.

- *Ontological authenticity* is defined as the extent to which individual respondents' (and the inquirer's) early constructions are improved, matured, expanded, and elaborated, so that all parties possess more information and become more sophisticated in its use. Useful *procedures* include the following: explication of the respondents' and the inquirer's a priori positions; comparison of respondents' earlier and later personal constructions; solicitation of respondents' and the inquirer's introspective statements about their own growth, as well as the testimony of selected respondents regarding their own changing constructions; and the establishment of an audit trail demonstrating changes.

- *Educative authenticity* is defined as the extent to which individual respondents (and the inquirer) possess enhanced understanding of, appreciation for, and tolerance of the constructions of others outside their own stakeholding group. Useful *procedures* include the following: employment of a peer debriefer and an

auditor by the inquirer, comparison of respondents' and the inquirer's assessments of the constructions held by others, respondents' and the inquirer's introspective statements about their understandings of others' constructions, respondent testimony, and maintenance of an audit trail.

- *Catalytic authenticity* is defined as the extent to which action (clarifying the focus at issue, moving to eliminate or ameliorate the problem, and/or sharpening values) is stimulated and facilitated by the inquiry process. Knowledge in and of itself is insufficient to deal with the constructions, problems, concerns, and issues that respondents bring to the inquiry process for elucidation; purposeful action must also be delineated. Useful *procedures* include the following: development of a joint construction (aiming at consensus when possible or explication of conflicting values when consensus is not possible), including the assignment of responsibility and authority for action; plans for respondent-inquirer collaboration; accessibility of the final report; and evidence of practical applications. Testimony of participants, the actual resolution of at least some of the inquiry concerns, and systematic follow-up over time to assess the continuing value of outcomes are also helpful techniques.

- *Tactical authenticity* is defined as the degree to which participants are empowered to take the action(s) that the inquiry implies or proposes. Useful *procedures* include the following: negotiation of data to be collected, as well as their interpretation and reporting; maintenance of CONFIDENTIALITY; use of consent forms; member checking/validation; and prior agreements about power. The best indicator that this criterion is satisfied is participant testimony. The presence of follow-up activities is highly indicative. Finally, the degree of empowerment felt by stakeholders is crucial.

The present state of developing authenticity criteria leaves much to be desired. Perhaps the most significant accomplishment to date is simply their existence, a demonstration of the fact that it is possible to "think outside the box" of conventional quality assessments. It is imperative for the inquirer to have thought through just what paradigm he or she elects to follow and, having made that decision, to select quality criteria that are appropriate. Applying positivist criteria to alternative paradigm inquiries (or conversely) can lead only to confusion and conflict.

—Egon G. Guba

REFERENCES

- Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology Journal*, 29, 75–92.
- Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

AUTOBIOGRAPHY

Autobiography refers to the telling and documenting of one's own life. Together with biography (researching and documenting the lives of others), autobiography has increasingly been drawn upon as a resource and method for investigating social life. Autobiographical research is part of a more general biographical turn within the social sciences, characterized by an emphasis on personal narratives and the LIFE HISTORY METHOD.

HISTORY OF AUTOBIOGRAPHICAL WORK

Autobiographical and life history work has a long genealogy, both generally and within the social sciences. Ken Plummer (2001) traces the rise of the “autobiographical society” and the various shifts that have taken place. He documents the ways in which the telling of lives has become inscribed within texts, with a move from oral to written traditions. He explores the possible origins of the autobiographical form and the rise of the individual “voice.” He also considers the ways in which individual life stories can become part of collective explorations of shared lives and experiences. Liz Stanley (1992) has traced the role of autobiography within social science (particularly sociology) and FEMINIST RESEARCH. She examines the methodological aspects of autobiographical work, highlighting the relationships between feminist praxis and autobiographical practice. Feminism has had a particular influence on the role of autobiographies and personal narratives within contemporary social science, both as a way of “giving voice” and as an approach that places an emphasis on self-reflexivity and personal knowledge. Autobiography, and life history work more generally, can create textual and discursive spaces for otherwise hidden or muted voices, such as marginal or oppressed social groups.

AUTOBIOGRAPHICAL RESOURCES AND NARRATIVES

Autobiographical data can be treated as resources by social scientists. General features of lives and experiences can be revealed through individual accounts and life stories. Written autobiographies can be considered as a rich data set of “lives” to be explored and analyzed in their own right, in terms of what they can reveal about a life, setting, organization, culture, event, or moment in time. Hence, stories of lives can be treated as realist accounts of social life. Autobiographical work is also composed and then articulated through written texts. Hence, autobiographies can be considered as research events—the processes through which lives are remembered, reconstructed, and written. There are collective conventions of memory and “ways” of (re)telling a life, marked by key events, places, people, and times. Autobiographical texts thus have particular genres and narrative forms that can be systematically explored. They are stories that draw on a standard set of literary and textual conventions. These can be analyzed for structure and form through, for example, NARRATIVE ANALYSIS.

AUTOBIOGRAPHY AND METHOD(OLOGY)

The “methods” of autobiographical work can also be used to gather social science data. Although autobiography is not a social science method in its own right, it does draw on a range of empirical materials or “documents of life” (Plummer, 2001). These include letters, DIARIES and other documentary sources, artifacts and biographical objects, and visual media such as film, video, and photography. Many of these sources of data are relatively underused within social science research, despite their analytical potential. Hence, a consideration of how “lives” are remembered, invoked, constructed, and reproduced can encourage a more eclectic approach to what might be considered as data about the social world.

In a consideration of feminist social science, Stanley (1992) suggests that work on autobiography (and biography) can be conceptualized as part of a distinctive methodological approach. Included within this is a focus on REFLEXIVITY; the rejection of conventional dichotomies such as objectivity-subjectivity, self-other, and public-private; and an emphasis on researching and theorizing experience. Thus, (auto)biographical approaches to social science

research can imply a greater self-awareness of the research process, research relationships, and the researcher-self, as well as a clearer appreciation of the value of LIVED EXPERIENCE and personal knowledge as part of social science scholarship.

WRITING AND RESEARCHING THE SELF

Personal narratives of the research process are a kind of autobiographical practice in their own right. The social researcher, as well as serving as a biographer of other lives, is simultaneously involved in autobiographical work of his or her own. Those working within QUALITATIVE RESEARCH, in particular, have reflected on and written about the self in a variety of autobiographical texts. FIELDNOTES and research journals have long been used to record the emotions, experiences, and identity work associated with social research. The genre of the confessional tale—revealing some of the personal aspects of fieldwork—allows for self-revelation and indiscretion, within a recognizable autobiographical format. Although such accounts are usually situated as separate from the research “data,” it is also possible to blend autobiography with analyses. For example, the “story” of the Japanese workplace told in *Crafting Selves* (Kondo, 1990) is presented as a multilayered text that challenges the dichotomies between researcher and researched, self and other. Kondo’s life stories form part of the representation of the research. She reveals herself and her biographical work as integral to the research process and product.

The self can also be a topic for explicit investigation. Thus, the methods of social research can be used to explicitly undertake autobiographical work. For example, ethnographic methods have been used to explore the personal experiences and life events of researchers themselves. These have included autobiographical analyses of illnesses, relationships, and bereavements. These personal autobiographical narratives can be located within a broader genre of AUTOETHNOGRAPHY (Reed-Danahay, 2001).

EVALUATION

It is difficult to evaluate autobiography as a particular approach to social research, given the variety of ways in which it may be used. Like all research that deals with life stories, autobiographical research

should be reflexive of the voices that are being heard and the lives that are being told. Storytelling voices are differentiated and stratified, and although we should not privilege the powerful, we should also not valorize the powerless in our treatment of their life stories.

A criticism that has been leveled at particular forms of autobiographical social research is the potential for romanticizing the self or of engaging in self-indulgence. Although distinctions can be made between social research and autobiographical practice, their intertextuality should also be recognized. The self is inexplicably linked to the processes of social research. This should be recognized and understood. Personal experiences and autobiographical stories can be sources for insightful analysis and innovative social science.

—Amanda Coffey

REFERENCES

- Kondo, D. K. (1990). *Crafting selves: Power, gender and discourses of identity in a Japanese workplace*. Chicago: University of Chicago Press.
- Plummer, K. (2001). *Documents of life 2*. London: Sage.
- Reed-Danahay, D. (2001). Autobiography, intimacy and ethnography. In P. Atkinson, A. Coffey, S. Delamont, J. Lofland, & L. Lofland (Eds.), *Handbook of ethnography* (pp. 405–425). London: Sage.
- Stanley, L. (1992). *The auto/biographical I: Theory and practice of feminist auto/biography*. Manchester, UK: Manchester University Press.

AUTOCORRELATION. See SERIAL CORRELATION

AUTOETHNOGRAPHY

Autoethnographic inquirers use their own experiences to garner insights into the larger culture or subculture of which they are a part. Autoethnography combines an autobiographical genre of writing with ethnographic research methods, integrated through both personal and cultural lenses and written creatively.

ETHNOGRAPHY first emerged as a method for studying and understanding the OTHER.

Fascination with exotic otherness attracted Europeans to study the peoples of Africa, Asia, the South Sea Islands, and the Americas. In the United States, for educated, White, university-based Americans, the others were Blacks, American Indians, recent immigrants, and the inner-city poor. To the extent that ethnographers reported on their own experiences as PARTICIPANT OBSERVERS, it was primarily through methodological reporting related to how they collected data and how, or the extent to which, they maintained detachment.

In contrast, an autoethnographic inquiry asks the following: How does my own experience of my own culture offer insights about my culture, situation, event, and/or way of life? Autoethnography raises the question of how an ethnographer might study her or his own culture without the burden or pretense of detachment (Patton, 2002). What if there is no *other* as the focus of study, but I want to study the culture of my own group, my own community, my own organization, and the way of life of people like me, the people I regularly encounter, or my own cultural experiences?

David Hayano (1979) is credited with originating the term *autoethnography* to describe studies by anthropologists of their own cultures. Autoethnographies are typically written in first-person voice but can also take the forms of short stories, poetry, fiction, photographic portrayals, personal essays, journal entries, and social science prose. Although autoethnographers use their own experiences to garner insights into the larger culture or subculture of which they are a part, great variability exists in the extent to which autoethnographers make themselves the focus of the analysis; how much they keep their role as social scientist in the foreground; the extent to which they use the sensitizing notion of culture, at least explicitly, to guide their analysis; and how personal the writing is. At the center, however, what distinguishes autoethnography from ethnography is self-awareness about and reporting of one's own experiences and introspections as a primary data source. Carolyn Ellis (Ellis & Bochner, 2000) has described this process as follows:

I start with my personal life. I pay attention to my physical feelings, thoughts, and emotions. I use what I call systematic sociological introspection and emotional recall to try to understand an experience I've lived through. Then I write my experience as a story. By exploring a particular life, I hope to understand a way of life. (p. 737)

BOUNDARY-BLURRING INQUIRY

Many social science academics object to the way autoethnography blurs the lines between social science and literary writing. Laurel Richardson (2000), in contrast, sees the integration of art, literature, and social science as precisely the point, bringing together creative and critical aspects of inquiry. She suggests that what these various new approaches and emphases share is that "they are produced through *creative analytic practices*," which leads her to call "this class of ethnographies *creative analytic practice ethnography*" (p. 929). Other terms are also being used to describe these boundary-blurring, evocative forms of qualitative inquiry—for example, *anthropological poetics*, *ethnographic poetics*, *ethnographic memoir*, *ethnobiography*, *evocative narratives*, *experimental ethnography*, *indigenous ethnography*, *literary ethnography*, *native ethnography*, *personal ethnography*, *postmodern ethnography*, *reflexive ethnography*, *self-ethnography*, *socioautobiography*, and *sociopoetics* (Patton, 2002, p. 85).

But how is one to judge the quality of such nontraditional social scientific approaches that encourage personal and creative ethnographic writing? In addition to the usual criteria of substantive contribution and methodological rigor, autoethnography and other evocative forms of inquiry add aesthetic qualities, REFLEXIVITY (consciousness about one's point of view and what has influenced it), impact on the reader, and authenticity of voice as criteria. In this vein, Elliot Eisner (1997) has suggested that in "the new frontier in qualitative research methodology," an artistic qualitative social science contribution can be assessed by the "number and quality of the questions that the work raises" as much as by any conclusions offered (p. 268).

Autoethnography can take the form of creative nonfiction. Sociologist Michael Quinn Patton (1999), for example, wrote about a 10-day Grand Canyon hike with his son, during which they explored what it means to *come of age*, or be initiated into adulthood, in modern society. To make the study work as a story and make scattered interactions coherent, he rewrote conversations that took place over several days into a single evening's dialogue, he reordered the sequence of some conversations to enhance the plot line, and he revealed his emotions, foibles, doubts, weaknesses, and uncertainties as part of the data of inquiry.

Autoethnography integrates ethnography with personal story, a specifically autobiographical

manifestation of a more general turn to BIOGRAPHICAL METHODS in social science that strives to “link macro and micro levels of analysis . . . [and] provide a sophisticated stock of interpretive procedures for relating the personal and the social” (Chamberlayne, Bornat, & Wengraf, 2000, pp. 2–3). Although controversial, the blurring of boundaries between social science and art is opening up new forms of inquiry and creative, evocative approaches to reporting findings.

—Michael Quinn Patton

REFERENCES

- Chamberlayne, P., Bornat, J., & Wengraf, T. (2000). *The turn to biographical methods in social science*. London: Routledge.
- Eisner, E. W. (1997). The new frontier in qualitative research methodology. *Qualitative Inquiry*, 3, 259–273.
- Ellis, C., & Bochner, A. P. (2000). Autoethnography, personal narrative, reflexivity: Researcher as subject. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 733–768). Thousand Oaks, CA: Sage.
- Hayano, D. M. (1979). Autoethnography: Paradigms, problems, and prospects. *Human Organization*, 38, 113–120.
- Patton, M. Q. (1999). *Grand Canyon celebration: A father-son journey of discovery*. Amherst, NY: Prometheus.
- Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks, CA: Sage.
- Richardson, L. (2000). Writing: A method of inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 923–948). Thousand Oaks, CA: Sage.

AUTOREGRESSION. See SERIAL CORRELATION

AVERAGE

The *average* is a term used to represent typical values of variables, particularly those values that indicate the midpoint of a distribution, the MEAN or the MEDIAN. In general use, it is taken to refer to the arithmetic mean. The arithmetic mean is the sum of the values of the given variable, divided by the number of cases. The median is the midpoint of the distribution of values, such that exactly half the values are higher and half are lower than the median. The choice of which

average to use depends on the particular advantages and disadvantages of the two measures for different situations and is discussed further under MEASURES OF CENTRAL TENDENCY.

Example

Most students, however innumerate, are able to calculate their grade point averages because it is typically the arithmetic mean of all their work over a course that provides their final grade. Table 1 shows a hypothetical set of results for students for three pieces of coursework. The table gives the mean and median for each student and also for the distribution of marks across each piece of work. It shows the different relationships the mean and median have with the different distributions, as well as the different ways that students can achieve a mark above or below the average. Dudley achieved a good result by consistent good marks, whereas Jenny achieved the same result with a much greater variation in her performance. On the other hand, Jenny achieved an exceptional result for Project 1, achieving 17 points above the mean and 19 above the median, whereas all students did similarly well on Project 2. If the apparently greater difficulty of Project 1 resulted in differential WEIGHTING of the scores, giving a higher weight to Project 1, then Jenny might have fared better than Dudley. The table provides a simple example of how the median can vary in relation to the mean. Of course, what would be the *fairest* outcome is a continuous matter of debate in educational institutions.

HISTORICAL DEVELOPMENT

The use of combining repeat observations to provide a check on their accuracy has been dated to Tycho Brahe's astronomical observations of the 16th century. This application of the arithmetic mean as a form of verification continued in the field of astronomy, but the development of the average for aggregation and estimation in the 18th and 19th centuries was intimately connected with philosophical debates around the meaning of nature and the place of mankind. Quetelet, the pioneer of analysis of social scientific data, developed the notion of the “average man” based on his measurements and used this as an ideal reference point for comparing other men and women. Galton, who owed much to Quetelet, was concerned rather with *deviation* from the average, with his stress

Table 1 Student Coursework Marks and Averages for Nine Students (Constructed Data)

	<i>Project 1</i>	<i>Project 2</i>	<i>Project 3</i>	<i>Coursework Average</i>	
				<i>Mean</i>	<i>Median</i>
Adam	52	62	63	59	62
Akiko	51	63	55	56.3	55
Chris	55	65	57	59	57
Dudley	63	62	64	63	62
Jenny	71	66	52	63	66
Maria	48	64	65	59	64
Robin	41	61	47	49.7	47
Surinder	47	67	60	58	60
Tim	58	66	42	55.3	58
Mean	54	64	56	58	59
Median	52	64	57	59	60

on “hereditary genius” and eugenics, from which he derived his understanding of regression. Since then, “average” qualities have tended to be associated more with mediocrity than with perfection.

AVERAGES AND LOW-INCOME MEASUREMENT

A common use of averages is to summarize income and income distributions. Fractions of average resources are commonly used as de facto poverty lines, especially in countries that lack an official poverty line. Which fraction, which average, and which measure of resources (e.g., income or expenditure) varies with the source and the country and with those carrying out the measurement. Thus, in Britain, for example, the annual *Households Below Average Income* series operationalizes what has become a semi-official definition of *low income* as 50% of mean income. The series does also, however, quote results for other fractions of the mean and for fractions of the median. By comparison, low-income information for the European Union uses 60% of the median as the threshold, and it is the median of expenditure that is being considered. In both cases, incomes have been weighted according to household size by equivalence scales. As has been pointed out, the fact that income distributions are skewed to the right means that the mean will come substantially higher up the income distribution than the median. It is therefore often considered that the median better summarizes

the center of an income distribution. However, when fractions of the average are being employed, the issue changes somewhat, both conceptually and practically. Conceptually, a low-income measure is attempting to grasp a point at which people are divorced from participation in expected or widespread standards and practices of living. Whether this is best expressed in relation to the median or the mean will depend in part on how the relationship of participation and exclusion is constructed. Similarly, at a practical level, a low-income measure is attempting to quantify the bottom end of the income distribution. Whether this is best achieved with an upper limit related to the mean or the median is a different question from whether the mean and the median themselves are the better summaries of the income distribution. In practice, in income distributions, 50% of the mean and 60% of the median typically fall at a very similar point. These two measures are consequently those that are most commonly used as they become effectively interchangeable.

An example of an income distribution is given in Figure 1. The figure shows gross U.S. household incomes from a survey of the late 1980s. Incomes have been plotted by household frequency in a form of histogram, although the histogram has been truncated so that the most extreme values (the highest incomes) have been excluded. The double line furthest to the right marks the mean of the distribution (at \$2,447 per month), with the solid line to the left of it indicating where the median falls (\$2,102). The mode, at the highest peak of the distribution, is also marked with

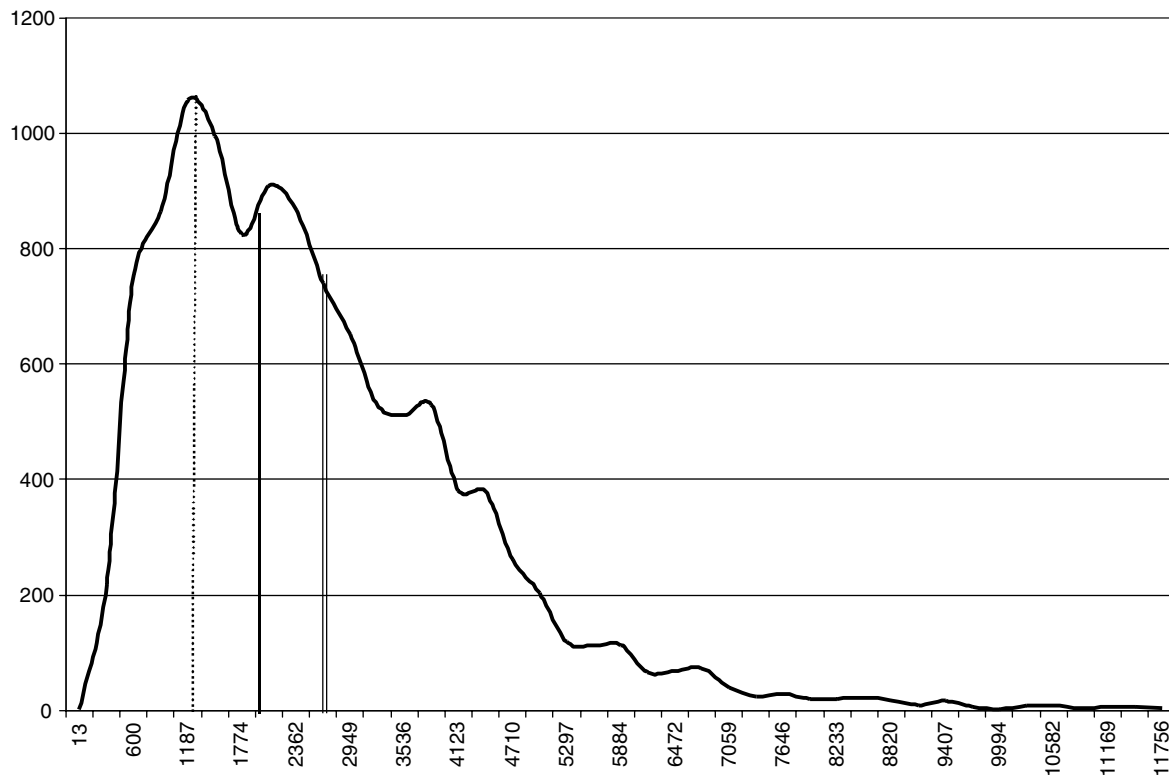


Figure 1 A Typical Income Distribution Showing Mode, Median, and Mean Income Points, U.S. Survey Data of Gross Household Incomes (Truncated), 1987

a dotted line. It is clear that the bulk of households fall to the left of the mean. The 60% of median and 50% of mean points will fall very close (at \$1,261 and \$1,223, respectively). They would thus cut off the bulk of the first hump of the distribution, which could then be deemed, on an inequality measure, to constitute low-income households.

—Lucinda Platt

See also MEAN, MEASURES OF CENTRAL TENDENCY, MEDIAN, MEDIAN TEST, *T*-TEST.

REFERENCES

- Goodman, A., Johnson, P., & Webb, S. (1997). *Inequality in the UK*. Oxford, UK: Oxford University Press.
- Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics: Vol. 1. Distribution theory* (5th ed.). London: Griffin.
- Yule, G. U., & Kendall, M. G. (1950). *An introduction to the theory of statistics* (14th ed.). London: Griffin.

B

BAR GRAPH

Bar graphs are one of the simplest ways of illustrating the relative frequencies or proportions of variables with discrete values. Typically, the bars representing the different values (e.g., age ranges, level of education, political affiliation, first language) rise from the x -axis, with the height of the bar representing the frequency, which can be read from the y -axis. It is also possible to produce bar graphs so that the bars are set along the y -axis and the frequencies are read from the x -axis, in which case the graph is called a horizontal bar graph. In this simple form, bar graphs display information that could also be displayed graphically using a two-dimensional PIE CHART. Figure 1 gives an example of a bar graph containing information on number of children in a group of 100 families. It gives an easy way of seeing the modal number of children (2) as well as the low frequency of larger families.

Figure 2 is an example of a horizontal bar graph that shows the proportions of different ethnic groups who made up the population of a British city according to the 1991 census.

Bar graphs can be adapted to contain information on more than one variable through the use of clustering or stacking. Figures 3 and 4 illustrate the information on ethnicity contained in Figure 2 when the age structure of the population is taken into account. Figure 3 is an example of a clustered bar chart where each column sums to 100%. This shows us that the various ethnic groups have different population structures, but does not tell us what that means in terms of absolute numbers. Figure 4, on the other hand, stacks

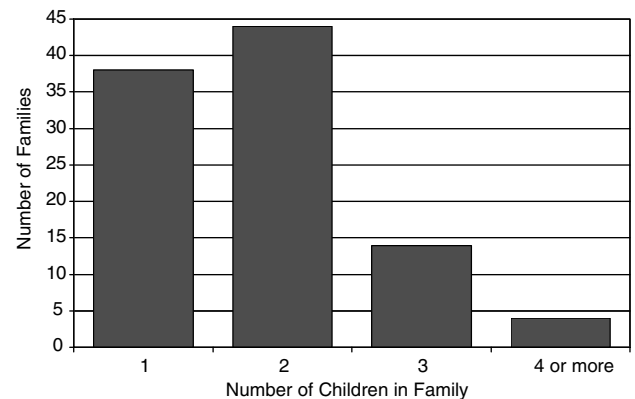


Figure 1 The Number of Children Per Family in a Sample of 100 Families

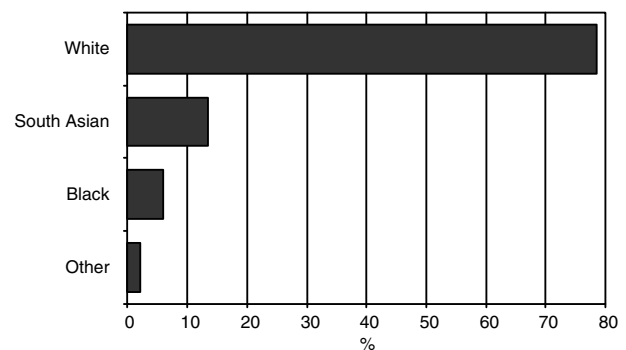


Figure 2 Bar Graph Showing the Ethnic Group Population Proportions in a British City, 1991

the populations next to each other, comparing white with all the minority groups, and retains the actual frequencies.

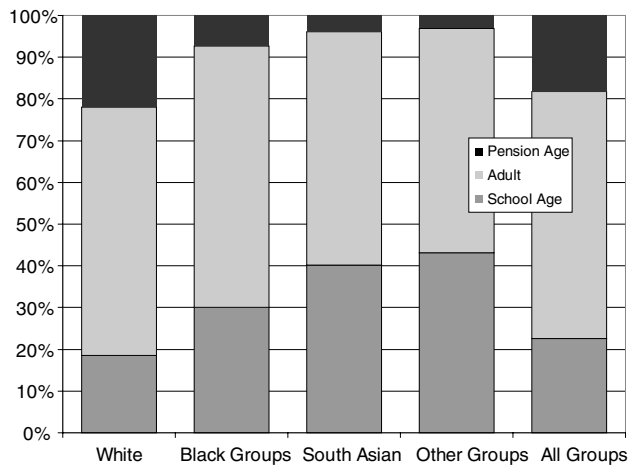


Figure 3 Stacked Bar Graph Showing the Population Structure of Different Ethnic Groups in a British City, 1991

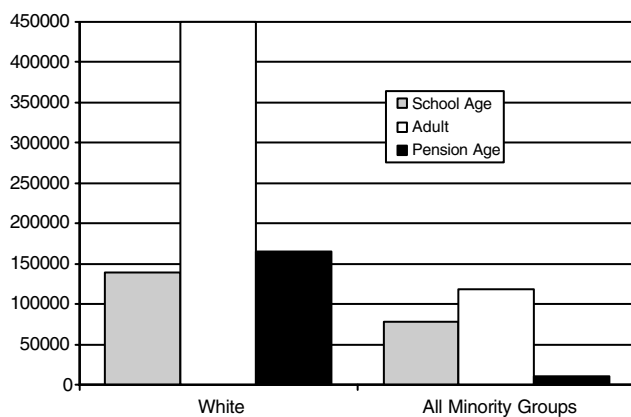


Figure 4 Clustered Bar Graph Comparing the Population Age Structure of White and Minority Ethnic Groups in a British City, 1991

The use of bar graphs to display data not only is a convenient way of summarizing information in an accessible form, but also, as John Tukey was at pains to argue, can lead us to insights into the data themselves. Therefore, it is perhaps surprising that the origins of the bar graph cannot be found before Charles Playfair's development of charts for statistical and social data at the end of the 18th century. Now, however, the development of software simplified the creation of graphs, including the more complex, three-dimensional stacked and clustered graphs, as well as the adjustment of their appearance, gridlines, and axes. The use of bar graphs in the social sciences to describe data and their characteristics swiftly and accessibly is

not only widespread but also expected. Nevertheless, the choices (e.g., type of bar graph, which data should be contained within the bar and which should form the axis, frequency or proportion) can still require considerable thought and may have implications for how the data are understood.

—Lucinda Platt

REFERENCES

- Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Wallgren, A. et al. (1996). *Graphing statistics and data: Creating better data*. Thousand Oaks, CA: Sage.

BASLINE

A baseline is a point of reference for comparison of research findings. For example, before implementation of a drug prevention program, researchers might establish the prevailing average drug use in the target population to serve as a baseline in evaluating the effects of the program. In a REGRESSION context, another example might be the expected value of DEPENDENT VARIABLE Y when INDEPENDENT VARIABLE $X = 0$; that value could stand as a baseline for comparing the expected value of Y when $X = 1$, or some other nonzero value. Sometimes, the results of an entire model may serve as the baseline, to be compared to another model.

—Michael S. Lewis-Beck

BASIC RESEARCH

Curiosity is the hallmark of human mental activity. Since the Greeks, it has been thought that a fundamental human impulse is to know the causes of things. This passion for knowledge has produced not only knowledge but also the foundations for vast improvements in our daily lives (Kornberg, 1997).

Basic research is driven by curiosity about the way the world works. We know that sooner or later, the new knowledge will have practical beneficial consequences, sometimes astounding ones. But these cannot

be foreseen, and, indeed, may coyly hide if searched for directly. The historical record makes abundantly clear that curiosity is the heartbeat of discovery and invention. Scientists engaged in basic research do not think about these things, except when called on to justify themselves; they are simply in thrall to their curiosities. But for those engaged in building the infrastructure of science or stocking the research portfolio, who often may be overpowered by concern for eradicating ills, the lessons about basic research are imperative (Kornberg, 1997).

Basic research can be characterized in several ways. Some note its tendency to be innovative and risk-taking. But at the bottom line, basic research is simply research whose purpose is to know the way some phenomenon or process works. Of course, phenomena and processes differ in the number and variety of domains they touch. For example, a high point in human understanding of physical nature was the realization that the same basic processes govern both earthly and celestial phenomena.

Thus, although curiosity drives all basic research, and all basic research seeks to understand the way something works, the more basic the process, the more basic the research. The Holy Grail is always discovery of the ultimate forces. Along the road, we lavish the utmost care and devotion on processes that, however ubiquitous or however long their reach, are not really ultimate forces. In so doing, we learn something about the way the world works, and we sharpen scientific tools in preparation for the more basic processes yet to be discerned.

Following Newton's (1686/1952) ideas for understanding physical nature, observed behavioral and social phenomena are usefully viewed as the product of the joint operation of several basic forces. Put differently, the sociobehavioral world is a *multi-factor* world, a view that lies at the heart of both theoretical and empirical work (e.g., Parsons, 1968). The scientific challenges are two—one theoretical, the other empirical. The theoretical challenge is to identify the basic forces governing human behavior, describe their operation, and derive their implications. The empirical challenge is to test the derived implications. The theoretical and empirical work jointly lead to the accumulation of reliable knowledge about human behavioral and social phenomena.

Even the briefest consideration of basic research would not be complete without at least some speculation on the identity of the basic forces governing

behavioral and social phenomena. Fermi used to spend an hour a day on pure speculation about the nature of the physical world (Segrè, 1970). In that spirit, an initial list of candidates for basic forces might include the four discussed in Jasso (2001, pp. 350–352):

1. To know the causes of things
2. To judge the goodness of things
3. To be perfect
4. To be free

—Guillermina Jasso

Further Reading

Kornberg's (1997) little essay should be required reading for all social scientists, all students, all deans, all review panels, and all R&D units.

REFERENCES

- Jasso, G. (2001). Rule-finding about rule-making: Comparison processes and the making of norms. In M. Hechter & K.-D. Opp (Eds.), *Social norms* (pp. 348–393). New York: Russell Sage.
- Kornberg, A. (1997). *Basic research, the lifeline of medicine*. Retrieved from <http://www.nobel.se/medicine/articles/research/>
- Newton, I. (1952). *Mathematical principles of natural philosophy*. Chicago: Britannica. (Originally published in 1686)
- Parsons, T. (1968). Émile Durkheim. In D. L. Sills (Ed.), *International encyclopedia of the social sciences*, Vol. 4. New York: Macmillan.
- Segrè, E. (1970). *Enrico Fermi: Physicist*. Chicago: University of Chicago Press.

BAYES FACTOR

Given data y and two models M_0 and M_1 , the Bayes factor

$$B_{10} = \frac{p(y | M_1)}{p(y | M_0)} = \left\{ \frac{p(M_1 | y)}{p(M_0 | y)} \right\} / \left\{ \frac{p(M_1)}{p(M_0)} \right\}$$

is a summary of the evidence for M_1 against M_0 provided by the data, and is also the ratio of posterior to prior odds (hence the name “*Bayes factor*”). Twice the logarithm of B_{10} is on the same scale as the deviance and likelihood ratio test statistics for model comparisons. Jeffreys (1961) suggests the following scale for interpreting the Bayes factor:

B_{10}	$2 \log B_{10}$	Evidence for M_1
< 1	< 0	Negative (support M_0)
1 to 3	0 to 2	Barely worth mentioning
3 to 12	2 to 5	Positive
12 to 150	5 to 10	Strong
> 150	> 10	Very strong

Note that B is a ratio of *marginal likelihoods*, that is, $p(y|M_k) = \int p(y|\theta_k, M_k) p(\theta_k) d\theta_k$, where θ_k is a vector of parameters of model M_k , $p(\theta_k)$ is the prior density of θ_k , and $p(y|\theta_k, M_k)$ is the likelihood of y under model M_k , $k \in \{0, 1\}$. Raftery (1996) provides a summary of methods of computing Bayes factors, and Good (1988) summarizes the history of the Bayes factor.

—Simon Jackman

See also BAYES' THEOREM

REFERENCES

- Good, I. J. (1988). The interface between statistics and philosophy of science. *Statistical Science*, 3, 386–397.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Clarendon.
- Raftery, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 163–187). London: Chapman and Hall.

BAYES' THEOREM, BAYES' RULE

At the heart of Bayesian statistics and decision theory is Bayes' Theorem, also frequently referred to as Bayes' Rule. In its simplest form, if H is a hypothesis and E is evidence, then the theorem is

$$\Pr(H|E) = \frac{\Pr(E \cap H)}{\Pr(E)} = \frac{\Pr(E|H)\Pr(H)}{\Pr(E)}$$

provided $\Pr(E) > 0$, so that $\Pr(H|E)$ is the PROBABILITY of belief in H after obtaining E , and $\Pr(H)$ is the *prior* probability of H before considering E . The left-hand side of the theorem, $\Pr(H|E)$ is usually referred to as the *posterior* probability of H . Thus, the theorem supplies a solution to the general problem of inference or induction (e.g., Hacking, 2001), giving us a mechanism for learning about a hypothesis H from data E .

Bayes' Theorem itself is uncontroversial: It is merely an accounting identity that follows from the axiomatic foundations of probability linking joint, conditional, and marginal probabilities, that is, $\Pr(A|B)\Pr(B) = \Pr(A \cap B)$, where $\Pr(B) \neq 0$. Thus, Bayes' Theorem is sometimes referred to as the *rule of inverse probability*, because it shows how a conditional probability B given A can be “inverted” to yield the conditional probability A given B (Leamer, 1978, p. 39). In addition, the theorem provides a way for considering two hypotheses, H_i and H_j , in terms of ODDS RATIOS; that is, it follows from Bayes' Theorem that

$$\frac{\Pr(H_i|E)}{\Pr(H_j|E)} = \frac{\Pr(E|H_i)\Pr(H_i)}{\Pr(E|H_j)\Pr(H_j)}$$

giving the posterior odds ratio for H_i over H_j (given E) equal to the *likelihood ratio* of E under H_i and H_j times the prior odds ratio. In the continuous case (e.g., learning about a real-valued parameter θ given data y), Bayes' Theorem is often expressed as

$$\pi(\theta|y) \propto f(y;\theta)\pi(\theta),$$

where $f(y;\theta)$ is the *likelihood function*, and the constant of proportionality $[f(y;\theta)\pi(\theta) d\theta]^{-1}$ does not depend on θ and ensures that the posterior density integrates to 1. This version of Bayes' Theorem is fundamental to Bayesian statistics, showing how the likelihood function (the probability of the data, given θ) can be turned into a probability statement about θ , given data y . Note also that the Bayesian approach treats the parameter θ as a random variable, and conditions on the data, whereas frequentist approaches consider θ a fixed (but unknown) property of a population from which we randomly sample data y .

The Reverend Thomas Bayes died in 1761. The result now known as Bayes' Theorem appeared in an essay attributed to Bayes, and it was communicated to the Royal Society after Bayes' death by Richard Price in 1763 (Bayes, 1763) and republished many times (e.g., Bayes, 1958). Bayes himself stated the result only for a uniform prior. According to Stigler (1986b), in 1774, Laplace (apparently unaware of Bayes' work) stated the theorem in its more general form (for discrete events). Additional historical detail can be found in Bernardo and Smith (1994, chap. 1), Lindley (2001), and Stigler (1986a, chap. 3).

—Simon Jackman

REFERENCES

- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53, 370–418.
- Bayes, T. (1958). An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45, 293–315.
- Bernardo, J., & Smith, A. F. M. (1994). *Bayesian theory*. Chichester, UK: Wiley.
- Hacking, I. (2001). *An introduction to probability and inductive logic*. Cambridge, UK: Cambridge University Press.
- Leamer, E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York: Wiley.
- Lindley, D. V. (2001). Thomas Bayes. In C. C. Heyde & E. Seneta (Eds.), *Statisticians of the centuries* (pp. 68–71). New York: Springer-Verlag.
- Stigler, S. (1986a). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap.
- Stigler, S. (1986b). Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1, 359–378.

BAYESIAN INFERENCE

The Bayesian statistical approach is based on updating information using a PROBABILITY theorem from the famous 1763 essay by the Reverend Thomas Bayes. He was an amateur mathematician whose work was found and published 2 years after his death by his friend Richard Price. The enduring association of an important branch of statistics with his name is due to the fundamental importance of this probability statement, now called Bayes' Law or BAYES' THEOREM, which relates conditional probabilities.

Start with two events of interest, X and Y , which are not independent. We know from the basic axioms of probability that the conditional probability of X given that Y has occurred is obtained by

$$p(X|Y) = \frac{p(X, Y)}{p(Y)},$$

where $p(X, Y)$ is “the probability that both X and Y occur,” and $p(Y)$ is the unconditional probability that Y occurs. The conditional expression thus gives the probability of X after some event Y occurs.

We can also define a different conditional probability in which X occurs first,

$$p(Y|X) = \frac{p(Y, X)}{p(X)}.$$

Because the joint probability that X and Y occur is the same as the joint probability that Y and X occur

$[p(X, Y) = p(Y, X)]$, then it is possible to make the following simple rearrangement:

$$\begin{aligned} p(X, Y) &= p(X|Y)p(Y), \\ p(Y, X) &= p(Y|X)p(X), \\ p(X|Y)p(Y) &= p(Y|X)p(X), \\ \therefore p(X|Y) &= \frac{p(X)}{p(Y)}p(Y|X) \end{aligned} \quad (1)$$

(for more details, see Bernardo and Smith, 1994, chap. 3). The last line is the famous Bayes' Law, and this is really a device for “inverting” conditional probabilities. It is clear that one could just as easily produce $p(Y|X)$ in the last line above by moving the unconditional probabilities to the left-hand side in the last equality. Bayes' Law is useful because we often know $p(X|Y)$ and would like to know $p(Y|X)$, or vice versa. The fact that $p(X|Y)$ is never equal to $p(Y|X)$ (that is, the probability that

$$\frac{p(X)}{p(Y)} = 1$$

is zero) is often called the *inverse probability problem*.

It turns out that this simple result is the foundation for the general paradigm of Bayesian statistics.

DESCRIPTION OF BAYESIAN INFERENCE

Bayesian inference is based on fundamentally different assumptions about data and parameters than are classical methods (Box & Tiao, 1973, chap. 1). In the Bayesian world, all quantities are divided into two groups: observed and unobserved. Observed quantities are typically the data and any known relevant constants. Unobserved quantities include PARAMETERS of interest to be estimated, MISSING DATA, and parameters of lesser interest that simply need to be accounted for. All observed quantities are fixed and are conditioned on. All unobserved quantities are assumed to possess distributional qualities and are therefore treated as random variables. Thus, parameters are now no longer treated as fixed unchanging in the total population, and all statements are made in probabilistic terms.

The inference process starts with assigning PRIOR DISTRIBUTIONS for the unknown parameters. These range from very informative descriptions of previous research in the field to deliberately vague and diffuse forms that reflect relatively high levels of ignorance. The prior distribution is not an inconvenience imposed

by the treatment of unknown quantities; instead, it is an opportunity to systematically include qualitative, narrative, and intuitive knowledge into the statistical model. The next step is to stipulate a likelihood function in the conventional manner by assigning a parametric form for the data and inserting the observed quantities. The final step is to produce a POSTERIOR DISTRIBUTION by multiplying the prior distribution and the likelihood function. Thus, the likelihood function uses the data to *update* the prior knowledge conditionally.

This process, as described, can be summarized by the simple mnemonic:

$$\begin{aligned} &\text{posterior probability} \propto \text{prior probability} \\ &\quad \times \text{likelihood function.} \end{aligned} \quad (2)$$

This is just a form of Bayes' Law where the denominator on the right-hand side has been ignored by using proportionality. The symbol " $\propto$ " stands for "proportional to," which means that constants have been left out that make the posterior sum or integrate to 1, as is required of probability mass functions and PROBABILITY DENSITY FUNCTIONS. Renormalizing to a proper form can always be done later, plus, using proportionality is more intuitive and usually reduces the calculation burden. What this "formula" above shows is that the posterior distribution is a compromise between the prior distribution, reflecting research beliefs, and the likelihood function, which is the contribution of the data at hand. To summarize, the Bayesian inference process is given in three general steps:

1. Specify a probability model that includes some prior knowledge about the parameters if available for unknown parameter values.
2. Update knowledge about the unknown parameters by conditioning this probability model on observed data.
3. Evaluate the fit of the model to the data and the sensitivity of the conclusions to the assumptions.

Nowhere in this process is there an artificial decision based on the assumption that some NULL HYPOTHESIS of no effect is true. Therefore, unlike the seriously flawed null hypothesis SIGNIFICANCE TEST, evidence is presented by simply summarizing this posterior distribution. This is usually done with quantiles and probability statements, such as the probability that the parameter of interest is less than/greater than some interesting constant, or the probability that this parameter occupies some region. Also note that if the posterior distribution is now treated as a new prior

distribution, it too can be updated if new data are observed. Thus, knowledge about the parameters of interest is updated and accumulated over time (Robert, 2001).

THE LIKELIHOOD FUNCTION

Suppose collected data are treated as a fixed quantity and we know the appropriate probability mass function or probability density function for describing the data generation process. Standard likelihood and Bayesian methods are similar in that they both start with these two suppositions and then develop estimates of the unknown parameters in the parametric model. MAXIMUM LIKELIHOOD ESTIMATION substitutes the unbounded notion of likelihood for the bounded definition of probability by starting with Bayes' Law:

$$p(\theta|\mathbf{X}) = \frac{p(\theta)}{p(\mathbf{X})} p(\mathbf{X}|\theta), \quad (3)$$

where θ is the unknown parameter of interest and $\mathbf{X}$ is the collected data. The key is to treat

$$\frac{p(\theta)}{p(\mathbf{X})}$$

as an unknown function of the data independent of $p(\mathbf{X}|\theta)$. This allows us to use $L(\theta|\mathbf{X}) \propto p(\mathbf{X}|\theta)$. Because the data are fixed, then different values of the likelihood function are obtained merely by inserting different values of the unknown parameter, θ .

The likelihood function, $L(\theta|\mathbf{X})$, is similar to the desired but unavailable inverse probability, $p(\mathbf{X}|\theta)$, in that it facilitates testing alternate values of θ to find a most probable value, $\hat{\theta}$. However, because the likelihood function is no longer bounded by 0 and 1, it is now important only *relative* to other likelihood functions based on differing values of θ . Note that the prior, $p(\theta)$, is essentially ignored here rather than overtly addressed. This is equivalent to assigning a uniform prior in a Bayesian context, an observation that has led some to consider classical inference to be a special case of Bayesianism: "Everybody is a Bayesian; some know it."

Interest is generally in obtaining the *maximum likelihood* estimate of θ . The value of the unconstrained and unknown parameter, θ , provides the maximum value of the likelihood function, $L(\theta|\mathbf{X})$. This value of θ , denoted $\hat{\theta}$, is the most likely to have generated the data given H_0 expressed through a specific parametric form relative to other possible values in the sample space of θ .

BAYESIAN THEORY

The Bayesian approach addresses the inverse probability problem by making distributional assumptions about the unconditional distribution of the parameter, θ , prior to observing the data, $\mathbf{X} : p(\theta)$. The prior and likelihood are joined with Bayes' Law:

$$\pi(\theta|\mathbf{X}) = \frac{p(\theta)L(\theta|\mathbf{X})}{\int_{\Theta} p(\theta)L(\theta|\mathbf{X})d\theta}, \quad (4)$$

where $\int_{\Theta} p(\theta)L(\theta|\mathbf{X})d\theta = p(\mathbf{X})$. Here, the $\pi(\cdot)$ notation is used to distinguish the posterior distribution for θ from the prior. The term in the denominator is generally not important in making inferences and can be recovered later by integration. This term is typically called the *normalizing constant*, the *normalizing factor*, or the *prior predictive distribution*, although it is actually just the marginal distribution of the data and ensures that $\pi(\theta|\mathbf{X})$ integrates to 1.

A more compact and useful form of (4) is developed by dropping this denominator and using proportional notation, because $p(\mathbf{X})$ does not depend on θ and therefore provides no relative inferential information about more likely values of θ :

$$\pi(\theta|\mathbf{X}) \propto p(\theta)L(\theta|\mathbf{X}), \quad (5)$$

meaning that the unnormalized *posterior* (sampling) distribution of the parameter of interest is proportional to the prior distribution times the likelihood function.

The maximum likelihood estimate is equal to the Bayesian posterior mode with the appropriate uniform prior, and they are asymptotically equal given *any* prior: Both are NORMALLY DISTRIBUTED in the limit. In many cases, the choice of a prior is not especially important because as the sample size increases, the likelihood progressively dominates the prior. Although the Bayesian assignment of a prior distribution for the unknown parameters can be seen as subjective (although *all* statistical models are actually subjective), there are often strong arguments for particular forms of the prior: Little or vague knowledge often justifies a diffuse or even uniform prior, certain probability models logically lead to particular forms of the prior (conjugacy), and the prior allows researchers to include additional information collected outside the current study.

SUMMARIZING BAYESIAN RESULTS

Bayesian researchers are generally not concerned with just getting a specific point estimate of the

parameter of interest, θ , as a way of providing empirical evidence in probability distributions. Rather, the focus is on describing the shape and characteristics of the posterior distribution of θ . Such descriptions are typically in the form of direct probability intervals (credible sets and highest posterior density regions), quantiles of the posterior, and probabilities of interest such as $p(\theta_l < 0)$.

Hypothesis testing can also be performed in the Bayesian setup. Suppose Θ_1 and Θ_2 represent two competing hypotheses about the state of some unknown parameter, θ , and form a partition of the sample space: $\Theta = \Theta_1 \cup \Theta_2$, $\Theta_1 \cap \Theta_2 = \phi$. Prior probabilities are assigned to each of the two outcomes: $\pi_1 = p(\theta \in \Theta_1)$ and $\pi_2 = p(\theta \in \Theta_2)$. This leads to competing posterior distributions from the two priors and the likelihood function: $p_1 = p(\theta \in \Theta_1|\mathbf{X})$ and $p_2 = p(\theta \in \Theta_2|\mathbf{X})$. It is common to define the prior odds, π_1/π_2 , and the posterior odds, p_1/p_2 , as evidence for H_1 versus H_2 . A much more useful quantity, however, is $(\pi_1/\pi_2)/(p_1/p_2)$, which is called the BAYES FACTOR. The Bayes Factor is usually interpreted as odds favoring H_1 versus H_2 given the observed data. For this reason, it leads naturally to the Bayesian analog of hypothesis testing.

AN EXAMPLE

Suppose that $x_1, x_2, \dots, x_n$ are observed independent random variables, all produced by the same Bernoulli probability mass function with parameter θ so that their sum,

$$y = \sum_{i=1}^n x_i,$$

is distributed binomial (n, θ) with known n and unknown θ . Assign a beta (A, B) prior distribution for θ ,

$$p(\theta|A, B) = \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} \theta^{A-1} (1-\theta)^{B-1},$$

$$0 < \theta; 0 < A \text{ or } B,$$

with researcher-specified values of A and B .

The posterior distribution for θ is obtained by the use of Bayes' Law:

$$\pi(\theta|y) = \frac{p(\theta|A, B)p(y|\theta)}{p(y)}, \quad (6)$$

where $p(y|\theta)$ is the probability mass function for y . The numerator of the right-hand side is easy to calculate now that there are assumed parametric forms:

$$\begin{aligned}
 p(y, \theta) &= p(y|\theta)p(\theta) \\
 &= \left[\binom{n}{y} \theta^y (1 - \theta)^{n-y} \right] \\
 &\quad \times \left[\frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} \theta^{A-1} (1 - \theta)^{B-1} \right] \quad (7) \\
 &= \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)} \\
 &\quad \times \theta^{y+A-1} (1 - \theta)^{n-y+B-1}
 \end{aligned}$$

and the denominator can be obtained by integrating θ out of this joint distribution:

$$\begin{aligned}
 p(y) &= \int_0^1 \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)} \\
 &\quad \times \theta^{y+A-1} (1 - \theta)^{n-y+B-1} d\theta \\
 &= \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)} \quad (8) \\
 &\quad \times \frac{\Gamma(y+A)\Gamma(n-y+B)}{\Gamma(n+A+B)}.
 \end{aligned}$$

For more technical details, see Gill (2002, p. 67). Performing arithmetic, the operations in (6) provide

$$\begin{aligned}
 \pi(\theta|y) &= \frac{p(\theta)p(y|\theta)}{p(y)} \\
 &= \frac{\Gamma(n+A+B)}{\Gamma(y+A)\Gamma(n-y+B)} \quad (9) \\
 &\quad \times \theta^{(y+A)-1} (1 - \theta)^{(n-y+B)-1}.
 \end{aligned}$$

This is a new beta distribution for θ with parameters $A' = y + A$ and $B' = n - y + B$ based on updating the prior distribution with the data y . Because the prior and the posterior have the same distributional family form, this is an interesting special case, and it is termed a *conjugate model* (the beta distribution is conjugate for the binomial probability mass function). Conjugate models are simple and elegant, but it is typically the case that realistic Bayesian specifications are much more complex analytically and may even be impossible.

MARKOV CHAIN MONTE CARLO

A persistent and troubling problem for those developing Bayesian models in the 20th century was that it was often possible to get a sufficiently complicated posterior from multiplying the prior and the likelihood, that the mathematical form existed, but quantities of interest such as means and quantiles could not be calculated analytically. This problem pushed Bayesian methods to the side of mainstream statistics for quite some time. What changed this unfortunate state of affairs was the publication of a review essay by Gelfand and Smith (1990) that described how similar problems had been solved in statistical physics with MARKOV CHAIN simulation techniques. Essentially, these are iterative techniques where the generated values are (eventually) from the posterior of interest (Gill, 2002). Thus, the difficult posterior form can be described empirically using a large number of simulated values, thus performing difficult integral calculations through simulation. The result of this development was a flood of papers that solved a large number of unresolved problems in Bayesian statistics, and the resulting effect on Bayesian statistics can be easily described as revolutionary.

Markov chains are composed of successive quantities that depend probabilistically only on the value of their immediate predecessor. In general, it is possible to set up a chain to estimate multidimensional probability structures, the desired posterior distributions, by starting a Markov chain in the appropriate sample space and letting it run until it settles into the desired target distribution. Then, when it runs for some time confined to this particular distribution, it is possible to collect summary statistics such as means, variances, and quantiles from the simulated values. The two most common procedures are the Metropolis-Hastings algorithm and the GIBBS SAMPLING, which have been shown to possess desirable theoretical properties that lead to empirical samples from the target distribution. These methods are reasonably straightforward to implement and are becoming increasingly popular in the social sciences.

—Jeff Gill

REFERENCES

- Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.
- Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. New York: Wiley.

- Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 389–409.
- Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. New York: Chapman and Hall.
- Robert, C. P. (2001). *The Bayesian choice: A decision theoretic motivation* (2nd ed.). New York: Springer-Verlag.

BAYESIAN SIMULATION. See MARKOV CHAIN MONTE CARLO METHODS

BEHAVIOR CODING

Behavior coding is a systematic, flexible, and cost-efficient method of examining the interaction between a survey interviewer and respondent. Coders listen to live or taped interviews and assign CODES for behaviors such as the accuracy with which interviewers read questions and the difficulties respondents exhibit in answering. Results of this coding can be used to monitor interviewer performance, identify problems with survey questions, or analyze the effects of methodological experimentation. Although each of these objectives requires somewhat different procedures, the general process is the same. A researcher must identify behavior to code; develop an appropriate coding scheme; select and train coders; assess reliability; supervise the coding task; and, finally, analyze the results.

Beginning in the late 1960s, survey methodologists conducted a variety of studies of interviews based on behavior coding of taped interviews (e.g., Cannell, Fowler, & Marquis, 1968). These studies provided valuable information about various components of the survey interview, including differences in behaviors in OPEN-ENDED and CLOSED-ENDED QUESTIONS, in characteristics of respondents, and in experienced and inexperienced interviewers. The most significant finding in these studies was that the surveys did not proceed as the researcher had expected. Questions were not administered as specified, nor were other techniques followed according to instructions. The first reaction was to blame the interviewer, attributing deficiencies in performance to poor training, inadequate supervision, or simply a lack of interviewer effort. More careful analysis, however, suggested that much of the difficulty

could be attributed to the techniques in which they were trained or inadequacies in the survey questions.

Not surprisingly, the first practical routine use of behavior coding in survey research was to monitor interviewer performance. Behavior coding is now used as a technique in the developmental stages of survey questions and is used routinely by many organizations in evaluating questions during PRETESTS (Oksenberg, Cannell, & Kalton, 1991). Most coding schemes developed for the purpose of evaluating questions examine the initial reading of the question by the interviewer, as well as all of the subsequent respondent behaviors. Interviewer reading codes evaluate the degree to which the question was read exactly as worded. Respondent behavior codes, at a minimum, attempt to capture behaviors related to comprehension difficulties (e.g., request for clarification); potential problems related to question wording or length (e.g., interruptions); and the adequacy of the response (e.g., inadequate response, qualified response, adequate response) as well as inability to respond (e.g., “don’t know” responses and refusals). High rates of “poor” behaviors are indicative of potential problems with a question wording or response options.

—Nancy Mathiowetz

See also INTERVIEWER EFFECTS, INTERVIEWER TRAINING, PRETESTING

REFERENCES

- Cannell, C., Fowler, J., & Marquis, K. (1968). *The influence of interviewer and respondent psychological and behavioral variables on the reporting of household interviews* (Vital and Health Statistics Series 2, No. 6). Washington, DC: National Center for Health Statistics.
- Oksenberg, L., Cannell, C., & Kalton, G. (1991). New strategies for pretesting survey questions. *Journal of Official Statistics*, 7(3), 349–365.

BEHAVIORAL SCIENCES

The behavioral sciences are focused on the explanation of human behavior. Psychology is the primary example, although other social sciences have behavioral subfields. For example, in political science, the study of voting behavior is well-developed, as is the study of consumer behavior in economics.

—Michael S. Lewis-Beck

BEHAVIORISM

Behaviorism, a term invented by John B. Watson (Watson, 1959), is the name of both a method and a metaphysics. On the grounds that first-person reports of one's private consciousness are suspect because they cannot be independently verified, the *methodological behaviorist* seeks to replace personal introspection of one's own states of mind with controlled observations of one's behavior by other people. Instead of taking an angry child's description of her emotion as definitive, the behaviorist would note whether she tries to damage the sorts of things that reportedly annoy her. On the grounds that we should regard as unreal what can be observed by only one person in only one way, the *metaphysical behaviorist* maintains that emotions and other so-called states of mind are not private possessions knowable only by the people and animals who have them, but publicly observable dispositions to behave in determinate ways. Thus, metaphysical behaviorists contend that, because anger normally *consists in* a disposition to damage things, we are observing a child's emotion by observing how it is manifest in her behavior. As thus defined, behaviorism is well caricatured by a joke its critics used to tell about two behaviorists who meet in the street. One says to the other, "You feel fine; how do I feel?"

The first-person method of introspecting what were called the phenomena of (human) consciousness dominated psychology during the 18th and 19th centuries. During most of the 20th century, the third-person method of observing behavior, both animal and human, under carefully controlled conditions dominated what was called behavioral science. However, by the end of the 20th century, many cognitive scientists became convinced that behaviorist strictures against introspection and consciousness had been *too* restrictive. In the new view, sensations, emotions, thoughts, and other states of mind are really states of the brain. These states will normally give rise to distinctive sorts of behavior, but they are related to the behavior as causes are related to their effects. So, it is argued, their true nature can no more be discovered merely by observing the behavior to which they give rise than the physical structure of magnets can be discovered merely by watching them attract iron filings. On the theory that the human brain has a special capacity to monitor its own states, some scientists and philosophers have therefore undertaken once again to study consciousness, the favorite

topic of 19th-century philosophers, by combining introspection with research in brain physiology. Behaviorists have responded to this renewal of an old program not by denying the relevance of brain physiology, but by insisting that the so-called phenomena of consciousness are illusions, not hard data, and that—as Howard Rachlin tries to show—behavior is still the key to the mind's secrets (Rachlin, 2000).

—Max Hocutt

REFERENCES

- Rachlin, H. (2000). *The science of self-control*. Cambridge, MA: Harvard University Press.
 Watson, J. B. (1959). *Behaviorism*. Chicago: University of Chicago Press.

BELL-SHAPED CURVE

The bell-shaped curve is the term used to describe the shape of a NORMAL DISTRIBUTION when it is plotted with the x -axis showing the different values in the distribution and the y -axis showing the frequency of their occurrence. The bell-shaped curve is a symmetric distribution such that the highest frequencies cluster around the midpoint of the distribution with a gradual tailing off toward 0 at an equal rate on either side in the frequency of values as they move away from the center of the distribution. In effect, it resembles a church bell, hence the name. Figure 1 illustrates such a bell-shaped curve. As can be seen from the symmetry and shape of the curve, all three MEASURES OF CENTRAL TENDENCY—the mean, the mode, and the median—coincide at the highest point of the curve.

The bell-shaped curve is described by its mean, μ , and its STANDARD DEVIATION, σ . Each bell-shaped curve with a particular μ and σ will represent a unique distribution. As the frequency of distributions is greater toward the middle of the curve and around the mean, the probability that any single observation from a bell-shaped distribution will fall near to the mean is much greater than that it will fall in one of the tails. As a result, we know, from normal probabilities, that in the bell-shaped curve, 68% of values will fall within 1 standard deviation of the mean, 95% will fall within roughly 2 standard deviations of

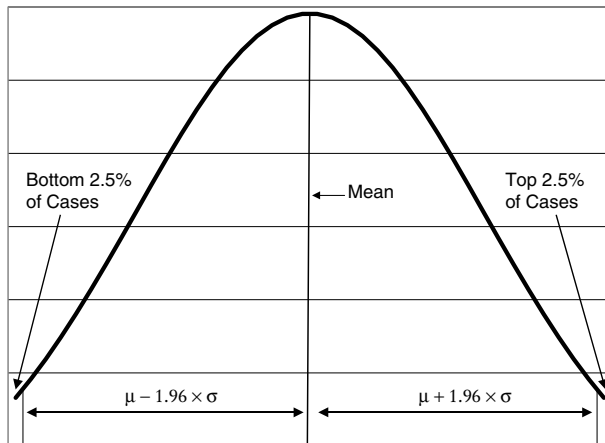


Figure 1 Example of a Bell-Shaped Curve

the mean, and nearly all will fall within 3 standard deviations of the mean. The remaining observations will be shared between the two tails of the distribution. This is illustrated in Figure 1, where we can see that 2.5% of cases fall into the tails beyond the range represented by $\mu \pm 1.96 \times \sigma$. This gives us the probability that in an approximately bell-shaped sampling distribution, any case will fall with 95% probability within this range, and thus, by statistical extrapolation, the population parameter can be predicted as falling within such a range with 95% confidence. (See also CENTRAL LIMIT THEOREM, NORMAL DISTRIBUTION.)

A particular form of the bell-shaped curve has its mean at 0 and a standard deviation of 1. This is known as the standard normal distribution, and for such a distribution the distribution probabilities represented by the equation $\mu \pm z^* \sigma$ simplify to the value of the multiple of σ (or Z-SCORE) itself. Thus, for a standard normal distribution, the 95% probabilities lie within the range ± 1.96 .

Bell-shaped distributions, then, clearly have particular qualities deriving from their symmetry, such that it is possible to make statistical inferences for any distributions that approximate this shape. By extrapolation, they also form the basis of statistical inference even when distributions are not bell-shaped. In addition, distributions that are not bell-shaped can often be transformed to create an approximately bell-shaped curve. Thus, for example, income distributions, which show a skew to the right, can be transformed into an approximately symmetrical distribution by taking the log of the values. Conversely, distributions with a skew to the left,

such as examination scores or life expectancy, can be transformed to approximate a bell shape by squaring or cubing the values.

HISTORICAL DEVELOPMENT AND USE

The normal-distribution, or bell-shaped, curve has held a particular attraction for social statisticians, especially since Pierre Laplace proved the central limit theorem in 1810. This showed how a bell-shaped curve could arise from a non-bell-shaped distribution, and how a set of samples, all with their own errors, could, through their means, take on a symmetric and predictable form. Laplace was building on the work of Carl Friedrich Gauss, who gave his name to the normal distribution (it is alternatively known as the Gaussian distribution). Astronomical observation, which employed the arithmetic MEAN as a way of checking repeat observations and providing the “true” value, also showed empirically that the other observations followed the shape of a bell-shaped curve, clustering around the mean. The bell curve therefore represented a set of errors clustering around a true value. The idea of the mean as representing some uniquely authentic value was taken up by Adolphe Quetelet in his work on the “average man” (see AVERAGE). Quetelet also developed the (erroneous) idea that normal distributions are the form to which aspects of social and biological life naturally tend and, conversely, assumed that where bell-shaped curves were found, the population creating the curve could be considered homogeneous. The tendency to regard bell-shaped distributions as having some innate validity and the temptation to reject those that do not continued to be a feature of statistical work into the 20th century, as Stigler (1999) has pointed out. However, although some distributions in nature (e.g., adult height) approximate a bell-shaped curve distribution in that they show a clustering around the mean and a tailing off in either direction, most do not; and perfect symmetry is rarely found. Nevertheless, the theoretical developments based on the bell-shaped curve and its properties, and the Gaussian distribution in particular, remain critical to statistical inference in the social sciences.

The visual and conceptual attraction of the bell-shaped curve has meant its continuing use as a marker for grasping truths or expressing beliefs about society beyond, although drawing on, its statistical

inferential properties. Thus, for example, Richard J. Herrnstein and Charles Murray's (1994) somewhat notorious discussion of the distribution of intelligence among whites and African Americans in the United States used the term *The Bell Curve* as the title for their book.

—Lucinda Platt

REFERENCES

- Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.
- Marsh, C. (1988). *Exploring data: An introduction to data analysis for social scientists*. Cambridge, UK: Polity.
- Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.
- Yule, G. U., & Kendall, M. G. (1950). *An introduction to the theory of statistics* (14th ed.). London: Charles Griffin.

BERNOULLI

Bernoulli is the name given to a PROBABILITY feature known as the Bernoulli process, used for describing the probabilities of an event with two outcomes, such as flipping a coin. The BINOMIAL probability of k successes in n trials of a Bernoulli process is given by

$$P(E) = \binom{n}{k} p^k (1-p)^{n-k},$$

where the event E has outcomes of exactly k heads during n flips of a fair coin.

—Tim Futing Liao

BEST LINEAR UNBIASED ESTIMATOR

The term *best linear unbiased estimator* (BLUE) comes from application of the general notion of unbiased and efficient estimation in the context of linear

estimation. In statistical and econometric research, we rarely have populations with which to work. We typically have one or a few SAMPLES drawn from a population. Our task is to make meaningful statistical inferences about parameter estimates based upon sample statistics used as estimators. Here, the statistical properties of estimators have a crucial importance because without estimators of good statistical properties, we could not make credible statistical inferences about population parameters. We desire estimators to be unbiased and efficient. In other words, we require the expected value of estimates produced by an estimator to be equal to the true value of population parameters. We also require the variance of the estimates produced by the estimator to be the smallest among all unbiased estimators. We call an estimator the *best unbiased estimator* (BUE) if it satisfies both conditions. The class of BUE estimators may be either linear or nonlinear. If a BUE estimator comes from a linear function, then we call it a best linear unbiased estimator (BLUE).

For a more sophisticated understanding of the term BLUE, we need to delve into the meaning of the terms “estimator,” “unbiased,” “best,” and “linear” in more detail.

ESTIMATOR

An estimator is a decision rule for finding the value of a population parameter using a sample statistic. For any population parameter, whether it is a population average, dispersion, or MEASURE OF ASSOCIATION between variables, there are usually several possible estimators. Good estimators are those that are, on average, close to the population parameter with a high degree of certainty.

UNBIASEDNESS

With unbiasedness, we require that, under repeated sampling, the expected value of the sample estimates produced by an estimator will equal the true value of the population parameter. For example, suppose we estimate from a sample a bivariate LINEAR REGRESSION $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i$ in order to find the parameters of a population regression function $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Unbiasedness requires that the expected values of $\hat{\beta}_0$ and $\hat{\beta}_1$ equal the actual values of β_0 and β_1 .

in the population. Mathematically, this is expressed $E[\hat{\beta}_0] = \beta$ and $E[\hat{\beta}_1] = \beta_1$. Intuitively, unbiasedness means that the sample estimates will, on average, produce the true population parameters. However, this does not mean that any particular estimate from a sample will be correct. Rather, it means that the estimation rule used to produce the estimates will, on average, produce the true population parameter.

EFFICIENCY

The term *best* refers to the efficiency criterion. Efficiency means that an unbiased estimator has minimum variance compared to all other unbiased estimators. For example, suppose we have a sample drawn from a normal population. We can estimate the midpoint of the population using either the mean or the median. Both are unbiased estimators of the midpoint. However, the mean is the more efficient estimator because, under repeated sampling, the variance of the sample medians is approximately 1.57 times larger than the variance of the sample means. The rationale behind the efficiency criterion is to increase the probability that sample estimates fall close to the true value of the population parameter. Under repeated sampling, the estimator with the smaller variance has a higher probability of yielding estimates that are closer to the true value.

LINEARITY

For an estimator to be linear, we require that predicted values of the dependent variable be a linear function of the estimated parameters. Notice that we are interested here in linearity in parameters, not in variables. For example, suppose we estimate a sample REGRESSION such as $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1\sqrt{x_i} + \varepsilon_i$. In this regression, the relation between $\hat{y}_i$ and x_i is not linear, but the relation between $\hat{y}_i$ and $\sqrt{x_i}$ is linear. If we think of $\sqrt{x_i}$, instead of x_i , as our explanatory variable, then the linearity assumption is met. However, this cannot be applied equally for all estimators. For example, suppose we estimate a function like

$$\hat{y}_i = \hat{\beta}_0 + \frac{\sqrt{\hat{\beta}_1}}{\hat{\beta}_0} x_i + \varepsilon_i.$$

This is not a linear estimator, but requires nonlinear estimation techniques to find $\hat{\beta}_0$ and $\hat{\beta}_1$. An

estimator of this sort might be BUE, but would not be BLUE.

THE GAUSS-MARKOV THEOREM

BLUE implies meeting all three criteria discussed above. BLUE estimators are particularly convenient from a mathematical standpoint. A famous result in statistics is the Gauss-Markov theorem, which shows that the LEAST SQUARES estimator is the best linear unbiased estimator under certain assumptions. This means that when an estimation problem is linear in parameters, it is always possible to use a least squares estimator to estimate population parameters without bias and with minimum variance. This result is convenient mathematically because it is unnecessary to prove the statistical properties of least squares estimators. They always have good statistical properties as long as the Gauss-Markov assumptions are met.

LIMITS OF BLUE

Although the BLUE property is attractive, it does not imply that a particular estimator will always be the best estimator. For example, if the disturbances are not normal and have been generated by a “fat-tailed” distribution so that large disturbances occur often, then linear unbiased estimators will generally have larger variances than other estimators that take this into account. As a result, nonlinear or robust estimators may be more efficient. Additionally, linear estimators may not be appropriate in all estimation situations. For example, in estimating the variance of the population disturbance (σ_ε^2), quadratic estimators are more appropriate.

—B. Dan Wood and Sung Ho Park

REFERENCES

- Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.
- Greene, W. (2003). *Econometric analysis* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.
- Judge, G. G., Hill, R. C., Griffiths, W. E., Lutkepohl, H., & Lee, T. (1988). *Introduction to the theory and practice of econometrics*. New York: Wiley.

BETA

This Greek letter, β , refers to a REGRESSION slope in two different ways. In the first, it is the symbol for the slope value in the regression equation for the POPULATION (as opposed to the SAMPLE). In the second, it is the symbol for a standardized slope (expressing change in STANDARD DEVIATION units) and is sometimes called a “beta weight.” Usually, the context will make clear which of the two meanings of beta is intended.

—Michael S. Lewis-Beck

See also STANDARDIZED REGRESSION COEFFICIENT

BETWEEN-GROUP DIFFERENCES

A relationship exists between two variables if there is covariation between two variables; that is, if certain values of one variable tend to go together with certain values of the other variable. If people with more education tend to have higher incomes than people with less education, a relationship exists between education and income.

To study a possible relationship, we create groups of observations based on the independent variable in such a way that all elements that have the same value on the independent variable belong to the same group. The independent variable can be a NOMINAL VARIABLE, such as gender, where all the females are in one group and all the males in another group. Next, the two groups are compared on the dependent variable, say, income. If the two groups are found to be different, a relationship exists.

The independent variable can also be a QUANTITATIVE VARIABLE, such as years of completed school. Here, everyone with 8 years of education is in one group, everyone with 9 years of education is in the next group, and so on. Again, with income as the dependent variable, we compare the incomes in the various groups.

One way to compare the groups is to compute the value of a summary statistic, such as the mean, and then compare these means across the groups. We could prefer to compare the medians from the different groups, or we can create a graph such as BOXPLOTS and compare them. Using the means, the question becomes, How do we compare them? Also, how large are the differences

in the means in comparison with the spread of the values of the dependent variable?

With a nominal independent variable, the method most commonly used to compare the means of the groups is ANALYSIS OF VARIANCE; with a quantitative independent variable, REGRESSION analysis is used the most. In analysis of variance, we compare the mean incomes of the two genders to the overall mean. In regression, we find the predicted income values of each year of education and compare these values to the overall mean. Because these points lie on a line, we need not ask if the points are different; rather, does the regression line through these points have a nonzero slope, thereby showing that the predicted incomes are different? Had we run an analysis of variance on the education-income data, we would have found a smaller RESIDUAL SUM OF SQUARES and a better fit than for the regression analysis. But an analysis of variance model uses more DEGREES OF FREEDOM than the corresponding regression model, and a regression analysis would therefore typically have a smaller *P* VALUE.

In analysis of variance, we compare the group means to the overall mean, taking the sizes of the groups into account, by computing the BETWEEN-GROUP SUM OF SQUARES. The more the groups vary, the larger the sum of squares becomes. This sum is also referred to as the *between sum of squares*. Presumably, the means are different because the dependent variable has been affected by one or more independent variables. If gender is the one independent variable, the sum might be labeled *gender sum of squares* or *model sum of squares*.

The similar quantity in regression analysis is most often simply called the *regression sum of squares*. It takes into account how steep the regression line is and how spread out the observations are on the independent variable.

—Gudmund R. Iversen

REFERENCES

- Cobb, G. W. (1997). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Iversen, G. R., & Norpoth, H. (1987). *Analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series 07-001, 2nd ed.). Newbury Park, CA: Sage.
- Lewis-Beck, M. S. (1980). *Applied regression: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series 07-022). Beverly Hills, CA: Sage.

BETWEEN-GROUP SUM OF SQUARES

In principle, in the study of relationships between variables, the data are first divided into groups according to one or more independent variables. Within each group, we often compute a summary measure of the dependent variable, such as a mean, and then we compare these means.

If the means are different, then we have established that a relationship exists between the independent and the dependent variable. With two groups, the comparison is easy: We simply subtract one mean from the other to see if they are different. With more than two groups, the comparisons of the means become more extensive. We want to take into account how different the means are from each other as well as how large the groups are.

With an independent NOMINAL VARIABLE having k values, we do an ANALYSIS OF VARIANCE to determine how different the groups are. In such an analysis, for mathematical and historical reasons, we compute the following sum:

$$n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + \cdots + n_k(\bar{y}_k - \bar{y})^2,$$

where the n s are the number of observations in the groups, the y -bars with subscripts are the k group means, and the one mean without a subscript is the overall mean of the dependent variable. This is the between-group sum of squares. Sometimes, this is abbreviated to simply the *between sum of squares* (BSS) or even the *model sum of squares*. The magnitude of the sum clearly depends upon how different the group means are from each other and how large the groups are. With several independent variables, there is a between-group sum of squares for each variable, usually labeled with the name of the variable.

In and of itself, the magnitude of the between-group sum of squares does not tell much. But when it is divided by the total sum of squares, the fraction gives the proportion of the variation in the dependent variable accounted for by the independent variable. Also, when divided by its degrees of freedom ($k - 1$), the fraction becomes the numerator of the F value used to find the p VALUE used to determine whether to reject the null hypothesis of no group differences.

When the independent variable is a QUANTITATIVE VARIABLE, REGRESSION analysis is usually preferred over analysis of variance. A linear model assumes the population means of the dependent variable on a

straight line. Rather than a NULL HYPOTHESIS asking whether these means are equal, we need to look only at the slope of the regression line. A horizontal line with slope equal to zero implies equal means.

The regression sum of squares takes the place of the between-group sum of squares. Its major advantage is that its number of DEGREES OF FREEDOM is based only on the number of independent variables (1 in simple regression), not on the number of groups. However, if we also can do an analysis of variance on the same data, the BSS is typically somewhat smaller than the regression sum of squares, meaning that analysis of variance fits the data better than regression.

—Gudmund R. Iversen

See also SUM OF SQUARES

REFERENCES

- Iversen, G. R., & Norpoth, H. (1987). *Analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series 07-001, 2nd ed.). Newbury Park, CA: Sage.
- Lewis-Beck, M. S. (1980). *Applied regression: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series 07-022). Beverly Hills, CA: Sage.

BIAS

In common usage, the term *bias* means predisposition, prejudice, or distortion of the truth. Although the general idea is the same, in statistics, we use the term bias in connection with the relationship between a (sample) STATISTIC (also called an ESTIMATOR) and its true (PARAMETER) value in the POPULATION.

The difference between the EXPECTED VALUE of a statistic and the parameter value that the statistic is trying to estimate is known as bias. To see what all this means, one can proceed as follows.

Let X be a RANDOM VARIABLE from some PROBABILITY DENSITY FUNCTION (PDF) with the following parameters: a (population) mean value of μ_x and (population) variance of σ_x^2 . Suppose we draw a RANDOM SAMPLE of n observations from this population. Let

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

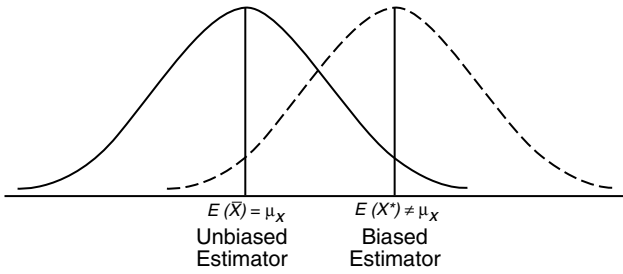


Figure 1 A Biased (X^*) and Unbiased ($\bar{X}$) Estimator of Population Mean Value, μ_x .

be the sample mean value. Suppose we use $\bar{X}$, the (sample) statistic, as an estimator of μ_x .

If the expected value of $\bar{X}$ is equal to μ_x , that is, $E(\bar{X}) = \mu_x$, then it can be said that $\bar{X}$ is an UNBIASED estimator of μ_x . If this equality does not hold, then $\bar{X}$ is a biased estimator of μ_x . That is,

$$\text{bias} = E(\bar{X}) - \mu_x,$$

which is, of course, zero if the expected value of the sample mean is equal to the true mean value. The bias can be positive or negative, depending on whether $E(\bar{X})$ is greater than μ_x or less than μ_x .

In other words, if one draws repeated samples of n observations of X , the average value of the means of those repeated samples should equal the true mean value μ_x . But note that there is no guarantee that any single value of $\bar{X}$ obtained from one sample will necessarily equal the true population mean value.

To find out if the expected value of the sample means equals the true mean value, one can proceed as follows:

$$\begin{aligned} E(\bar{X}) &= E\left[\frac{1}{n}(x_1 + x_2 + \cdots + x_n)\right] \\ &= \frac{1}{n}[Ex_1 + Ex_2 + \cdots + Ex_n] \\ &= \frac{1}{n}[n \cdot \mu_x] = \mu_x, \end{aligned}$$

because each X has the same mean value μ_x . Because $E(\bar{X}) = \mu_x$, it can be said that the sample mean is an unbiased estimator of the population mean.

Geometrically, the bias can be illustrated as in Figure 1.

To further illustrate bias, consider two estimators of the population variance σ_x^2 , namely,

$$S_x^{*2} = \sum \frac{(X_i - \bar{X})^2}{n} \quad \text{and} \quad S_x^2 = \sum \frac{(X_i - \bar{X})^2}{n-1},$$

which are the sample variances. The difference between the two is that the former has n DEGREES OF FREEDOM (df) and the latter has $(n-1) df$. It can be shown that

$$E(S_x^{*2}) = \sigma_x^2 \left(1 - \frac{1}{n}\right) \quad \text{but} \quad E(S_x^2) = \sigma_x^2.$$

Because the expected value of the former estimator is not equal to true σ_x^2 , it is a biased estimator of the true variance.

Unbiasedness is one of the desirable properties that an estimator should satisfy. The underlying idea is intuitively simple. If an estimator is used repeatedly, then, on average, it should be equal to its true value. If this does not happen, it will systematically distort the true value.

Unbiasedness is a small, or finite, sample property. An estimator may not be unbiased in small samples, but it may be unbiased in large samples, that is, as the sample size n increases indefinitely, the expected, or mean, value of the estimator approaches its true value. In that case, the estimator is called asymptotically unbiased. This is clearly the case with the estimator S_x^{*2} , because

$$\lim_{n \rightarrow \infty} E(S_x^{*2}) = \sigma_x^2.$$

Hence, asymptotically, S_x^{*2} is unbiased. In statistics, one comes across situations in which an estimator is biased in small samples, but the bias disappears as the sample size increases indefinitely.

The examples given thus far relate to a single random variable. However, the concept of bias can be easily extended to the multivariate case. For instance, consider the following MULTIPLE REGRESSION:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i,$$

where Y is the dependent variable, the X s are the explanatory variables, and u_i is the STOCHASTIC error term. In matrix notation, the preceding multiple regression can be written as

$$\mathbf{y} = \beta \mathbf{X} + \mathbf{u},$$

where $\mathbf{y}$ is an $(n \times 1)$ vector of observations on the dependent variable and $\mathbf{X}$ is an $(n \times k)$ data matrix on the k explanatory variables (including the intercept) and where $\mathbf{u}$ is an $(n \times 1)$ vector of the stochastic error terms.

Using the method of ORDINARY LEAST SQUARES (OLS), the estimators of the beta coefficients are obtained as follows:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

where $\hat{\beta}$ is a $(k \times 1)$ vector of the estimators of the true β coefficients.

Under the assumptions of the classical LINEAR REGRESSION model (the classical linear regression model assumes that values of the explanatory variables are fixed in repeated sampling, that the errors terms are not correlated, and that the error variance is constant, or HOMOSKEDASTIC), it can be shown that $\hat{\beta}$ is an unbiased estimator of the true β . The proof is as follows:

$$\begin{aligned} E(\hat{\beta}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u})] \\ &= \beta + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}] = \beta, \end{aligned}$$

where use is made of the assumptions of the classical linear regression model (note that the explanatory variables are nonstochastic).

As this expression shows, each element of the *estimated* β vector is equal to the corresponding value of the true β vector. In other words, each $\hat{\beta}_i$ is an unbiased estimator of the true β_i .

—Damodar N. Gujarati

REFERENCES

Gujarati, D. (2002). *Basic econometrics* (4th ed.). New York: McGraw-Hill.

BIC. See GOODNESS-OF-FIT MEASURES

BIMODAL

It is a characteristic of a statistical DISTRIBUTION having two separate peaks or MODES. In contrast, the NORMAL DISTRIBUTION is a UNIMODAL DISTRIBUTION.

—Tim Futing Liao

BINARY

Binary means two, and it may have at least three kinds of connotation to social scientists. The *binary system*, unlike the decimal system, has only two digits (0 and 1) and serves as the foundation of the operation of the modern computer. A *binary variable* measures a quantity that has only two categories. Social scientists often engage in *binary thinking* in categorizing social phenomena, such as the public and the private, Gemeinschaft and Gesellschaft, the micro and the macro, and so on.

—Tim Futing Liao

See also ATTRIBUTE, DICHOTOMOUS VARIABLES

BINOMIAL DISTRIBUTION

In a binomial distribution, there are a finite number of independently sampled observations, each of which may assume one of two outcomes. Thus, the binomial distribution is the DISTRIBUTION of a binary variable, such as males versus females, heads versus tails of a coin, or live versus dead seedlings. The SAMPLE proportion or PROBABILITY, p , that a given observation has a given outcome is calculated as the count, X , of observations sampled having that outcome divided by sample size, n . The sum of probabilities of each of a pair of binary states is always 1.0.

The binomial distribution is approximately what one would get if marbles were dropped at the apex of a Galton board, which is a triangular array of pegs such that there is a 50% chance of a dropped marble going either left or right. It is also the distribution resulting from the binomial expansion of the formula $(p + q)^n$, where p and q are the probabilities of getting or not getting a particular dichotomous outcome in n trials. For instance, (see Figure 1) for four trials selecting for Democrats and Republicans from a population with equal numbers of each, the binomial expansion is

$$\begin{aligned} (p + q)^4 &= (1/2 + 1/2)^4 \\ &= 1p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + 1q^4 \\ &= 1/16 + 4/16 + 6/16 + 4/16 + 1/16. \end{aligned}$$

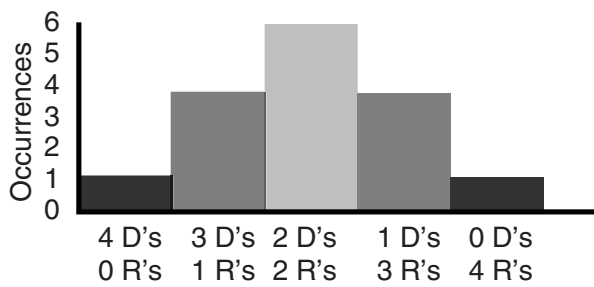


Figure 1 Binomial Distribution

The equation for calculating probabilities in a binomial distribution is

$$p(r) = {}_nC_r p^r q^{n-r},$$

where $p(r)$ refers to the probability of getting r observations in one category of a DICHOTOMOUS VARIABLE and $(n - r)$ observations in the other category, given sample size n ; p is the probability of getting the first state of the binary variable, and q is the probability of getting the other state; and ${}_nC_r$ is the binomial coefficient, which is the number of combinations of n things taken r at a time. When $p = q = .5$, the binomial distribution will be symmetrical, otherwise not.

For example, when tossing coins, the probability of getting either a head or a tail is .5. For three coin tosses, the probability of getting three heads is $(.5)(.5)(.5) = .125$. We could also get a head, a head, and a tail; or we could get a head, a tail, and a head. The number of combinations where we could get two heads in three tries is ${}_3C_2$, which is $3!/[2!(3 - 2)!]$, or 3. The probability of getting two heads is then

$$\begin{aligned} p(r) &= {}_nC_r p^r q^{n-r} = (3)(.5^2)(.5^{3-2}) \\ &= (3)(.25)(.5) = .375. \end{aligned}$$

A table of the binomial distribution, based on similar calculations, is printed in many statistics books or calculated by statistical software. For $n = 3$, the binomial table is

<i>r</i>	<i>p</i>		
	.1	.5	.9
0	0.729	0.125	0.001
1	0.243	0.375	0.027
2	0.027	0.375	0.243
3	0.001	0.125	0.729

Thus for $p = .1, .5$, or $.9$, the table above shows the probability of getting $r = 0, 1, 2$, or 3 heads or

other binomial states in three trials. Actual binomial tables typically show additional columns in p increments of .1, and also provide tables for $n = 2$ through $n = 20$. Some provide both one-sided and double-sided probabilities, where the former are appropriate to “as large/small” or “larger/smaller than” questions and the latter are appropriate to “as or more different from” questions.

When sample size n is large and the probability of occurrence is moderate, the sample proportion p and the count X will have a distribution that approximates the NORMAL DISTRIBUTION. The mean of this distribution will be approximately np , and its variance will be approximately $np(1 - p)$, which are the mean and variance of a binomial distribution. These normal approximations may be inaccurate for small values of n , so a common rule of thumb is that they may be used only if the product np is more than or equal to 10 and the product $np(1 - p)$ is also more than or equal to 10. For less than that, exact binomial probabilities are calculated. Statistical software usually will not use normal approximations to the binomial probability but will calculate exact binomial probabilities even for larger samples.

When sample size is large and the probability of occurrence is very small or large, the binomial distribution approaches the POISSON DISTRIBUTION. When samples are taken from a small POPULATION, the hypergeometric distribution may be used instead of the binomial distribution.

—G. David Garson

REFERENCES

Blalock, H. M. (1960). *Social statistics*. New York: McGraw-Hill.
Von Collani, E., & Drager, K. (2001). *Binomial distribution handbook*. Basel, Switzerland: Birkhauser.

BINOMIAL TEST

The binomial test is an exact probability test, based on the rules of PROBABILITY. It is used to examine the DISTRIBUTION of a single dichotomous variable when the researcher has a small SAMPLE. It is a non-parametric statistic that tests the difference between a sampled proportion and a given proportion, for one-sample tests.

The binomial test makes four basic assumptions. It assumes the variable of interest is a DICHOTOMOUS VARIABLE whose two values are mutually exclusive and exhaustive for all cases. Like all SIGNIFICANCE TESTS, the binomial test also assumes that the data have been sampled at random (see RANDOM SAMPLING). It assumes that observations are independently sampled. Finally, the binomial test assumes that the probability of any observation, p , has a given value that is constant across all observations. If an error-free value of p cannot be assumed, then alternative tests (Fisher's text, CHI-SQUARE TEST, t -TEST) are recommended. The binomial test is nonparametric and does not assume the data have a NORMAL DISTRIBUTION or any other particular distribution.

In a binomial test, we determine the probability of getting r observations in one category of a dichotomy and $(n - r)$ observations in the other category, given sample size n . Let p = the probability of getting the first category and let $q = (1 - p)$ = the probability of getting the other category. Let ${}_nC_r$ be the number of combinations of n things taken r at a time. The formula for the binomial probability is

$$p(r) = {}_nC_r p^r q^{n-r} = (n! p^r q^{n-r}) / [r!(n-r)!].$$

For example, assume we know that a particular city is 60% Democratic ($p = .60$) and 40% non-Democratic ($q = .40$), and we sample a particular fraternal organization in that city and find 15 Democrats ($r = 15$) in a sample of 20 people ($n = 20$). We may then ask if the 15:5 split found in our sample is significantly greater than the 60:40 split known to exist in the POPULATION. This question can be reformulated as asking, "What is the probability of getting a sample distribution as strong or stronger than the observed distribution?" The answer is the sum of the binomial probabilities $p(15), p(16), \dots, p(20)$. For instance, $p(15) = [(20!)(60^{15})(40^5)] / [(15!)(5!)] = .0746$. It can be seen that when n is at all large, computation requires a computer. Some statistics books also print a table of individual or cumulative binomial probabilities for various levels of n, r , and p .

The binomial mean is np and is the expected number of successes in n observations. The binomial standard deviation is the square root of npq , decreasing as p approaches 0 or 1.

When n is greater than 25 and p is near .50, and the product of npq is at least 9, then the binomial distribution approximates the normal distribution. In

this situation, a normal curve z -TEST may be used as an approximation of the binomial test. Note that the normal approximation is useful only when manual methods are necessary. SPSS, for instance, always computes the exact binomial test.

—G. David Garson

REFERENCES

- Conover, W. J. (1998). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.
- Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical inference*. New York: Wiley.
- Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

BIOGRAPHIC NARRATIVE INTERPRETIVE METHOD (BNIM)

This methodology for conducting and analyzing biographic narrative interviews has been used in a variety of European research projects, either directly (e.g., Chamberlayne, Rustin, & Wengraf, 2002; Rosenthal, 1998) or in a modified version (e.g., FREE ASSOCIATION INTERVIEWING). Assuming that "narrative expression" is expressive both of conscious concerns and also of unconscious cultural, societal, and individual pre-suppositions and processes, it is psychoanalytic *and* sociobiographic in approach.

In each BNIM interview, there are three subsessions (Wengraf, 2001, chap. 6). In the first, the interviewer offers only a carefully constructed single narrative question (e.g., "Please tell me the story of your life, all the events and experiences that have been important to you personally, from wherever you want to begin up to now").

In the second, sticking strictly to the sequence of topics raised and the words used, the interviewer asks for more narratives about them. A third subsession can follow in which nonnarrative questions can be posed.

The transcript thus obtained is then processed twice (Wengraf, 2001, Chap. 12). The strategy reconstructs the experiencing of the "interpreting and acting" subject as he or she interpreted events; lived his or her life; and, in the interview, told his or her story. This strategy requires the analysts to go forward through the events

as did the subject: *future-blind*, moment by moment, not knowing what comes next or later.

First, a chronology of objective life events is identified. "Objective life events" are those that could be checked using official documents, such as records of school and employment and other organizations. Each item (e.g., "failed exams at age 16"), stripped of the subject's current interpretation, is then presented separately to a research panel, which is asked to consider how this event might have been experienced *at the time*, and, if that experiential hypothesis were true, what might be expected to occur next or later in the life (following hypotheses). After these are collected and recorded, the next life-event item is presented: Its implications for the previous experiential and following hypotheses are considered, and a new round of hypothesizing commences. A process of imaginative identification and critical understanding is sought, constantly to be corrected and refined by the emergence of future events as they are presented one-by-one.

The transcript is then processed into "segments." A new segment is said to start when there is a change of speaker, of topic, or of the manner in which a topic is addressed. Again, each segment is presented in turn to a research panel that attempts to imagine how such interview events and actions might have been experienced at that moment of the interview, with subsequent correction and refinement by further segments.

In both series, separate structural hypotheses are sought, and only later are structural hypotheses developed relating the two. A similar future-blind procedure is also carried out for puzzling segments of the verbatim text (microanalysis). The question about the dynamics of the case can then be addressed: "Why did the people who lived their lives like *this*, tell their stories like *that*?"

—Tom Wengraf

REFERENCES

- Chamberlayne, P., Rustin, M., & Wengraf, T. (Eds.). (2002). *Biography and social exclusion in Europe: Experiences and life journeys*. Bristol, UK: Policy Press.
- Rosenthal, G. (Ed.). (1998). *The Holocaust in three generations: Families of victims and persecutors of the Nazi regime*. London: Cassell.
- Wengraf, T. (2001). *Qualitative research interviewing: Biographic-narrative and semi-structured method*. London: Sage.

BIOGRAPHICAL METHOD.

See LIFE HISTORY METHOD

BIPLOTS

The biplot is an essential tool for any analysis of a data matrix of subjects by variables in which not only the structure of the variables is of interest, but also differences between individuals. Surprisingly, in many large statistical software packages, it is not a standard tool. Several variants of the biplot exist, but only the basic is mentioned here. The biplot was probably first conceived by Tucker (1960), but Gabriel's (1971) seminal work first showed the full potential of the technique. The most extensive reference at present is Gower and Hand (1996).

A standard biplot is the display of a rectangular matrix $\mathbf{X}$ decomposed into a product $\mathbf{AB}'$ of an $I \times S$ matrix $\mathbf{A} = (a_{is})$ with row coordinates and a $J \times S$ matrix $\mathbf{B} = (b_{js})$ with column coordinates, where S indicates the number of coordinate axes, which, preferably, is two or three. Using this decomposition for $\hat{\mathbf{X}}$, a two-dimensional approximation of $\mathbf{X}$, each element $\hat{x}_{ij}$ of this matrix can be written as

$$\hat{x}_{ij} = a_{i1}b_{j1} + a_{i2}b_{j2},$$

which is the *inner* (or *scalar*) *product* of the row vectors (a_{i1}, a_{i2}) and (b_{j1}, b_{j2}) . A biplot is obtained by representing each row as a point a_i with coordinates (a_{i1}, a_{i2}) , and each column as point b_j with coordinates (b_{j1}, b_{j2}) , in a two-dimensional graph (with origin O). These points are generally referred to as *row markers* and *column markers*, respectively.

For example, suppose new mothers have been scored on variables that assess how well they are coping with their infants, such as parenting stress, depression, and cognitive stimulation of her child. The biplot of these data will display which mothers are coping in similar or in different ways, as expressed through their scores on the variables, and it shows which variables have similar patterns (say, depression and parenting stress) or different patterns of scores (depression and cognitive stimulation of the child).

In situations such as this, the variables typically will have been standardized and the coordinates of the

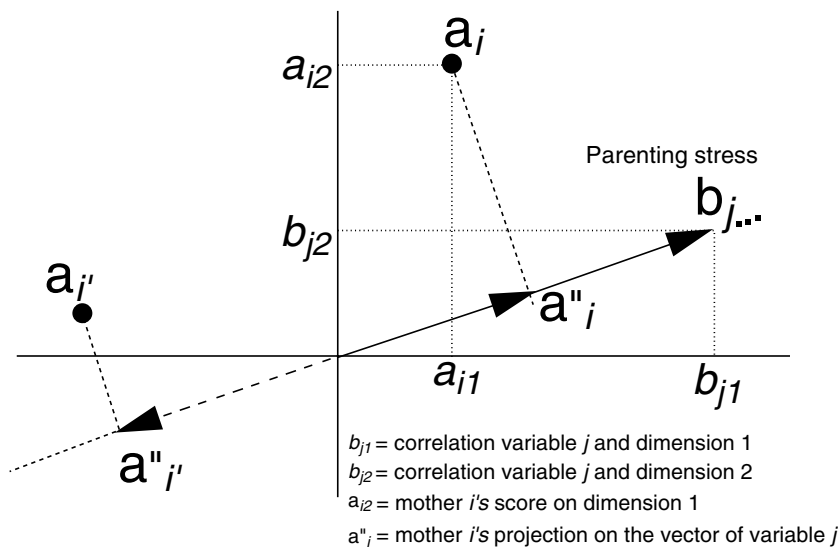


Figure 1 Main Elements of a Biplot: Column Vector, Row Points, Row Coordinates, Column Coordinates, and Projections

NOTE: The inner product between a_i and b_j is the product of the lengths of a''_i and b_j .

variables will have been scaled in such a way that these coordinates are the correlation between the variables and the coordinates (see also PRINCIPAL COMPONENTS ANALYSIS). In this case, the variables have *principal coordinates*, and the subjects have *standard coordinates*, that is, their coordinate axes have zero means and unit lengths. Because of this, the variables are usually displayed either as vectors originating from the origin, with the length of the vector indicating its variability, or as (biplot) axes on which the size of the original scale is marked, so-called calibrated axes. The subjects are indicated as points, and a subject's estimated score on a variable is deduced from the projection of its point on the variables, as illustrated in Figure 1. Because it is not easy to evaluate inner products in a three-dimensional space, the most commonly used biplots have only two dimensions and display the best two-dimensional approximation $\hat{\mathbf{X}}$ to the data matrix $\mathbf{X}$.

Given that a two-dimensional biplot is a reasonable approximation to the data matrix, the relative scores of two mothers on the same variable, say, parenting stress, can be judged by comparing the lengths of their projections onto the vector of parenting stress. In Figure 1 this entails comparing the length of a''_i with that of $a''_{i'}$. If a mother scores above the

average, her projected score will be on the positive side of the variables; if she scores below the average, her projected score will be on the negative side of the vector; when her projected score equals the average, it coincides with the origin of the biplot. Two variables having an acute (obtuse) angle with each other have a positive (negative) correlation. Absence of correlation results in perpendicular vectors. All of these statements are conditional on the biplot being a sufficiently good approximation to the data in $\mathbf{X}$.

An important point in constructing graphs for biplots is that the physical vertical and horizontal coordinate axes should have the same physical scale. This will ensure that when row points are projected onto a column vector, they

will end up in the correct place. Failing to adhere to this scaling (as is the case in some software packages) will make it impossible to evaluate inner products in the graph.

Besides being a valuable technique in its own right, the biplot also plays an important role in CORRESPONDENCE ANALYSIS OF CONTINGENCY TABLES and in three-mode data analysis.

—Pieter M. Kroonenberg

REFERENCES

- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58, 453–467.
- Gower, J. C., & Hand, D. J. (1996). *Biplots*. London: Chapman & Hall.
- Tucker, L. R. (1960). Intra-individual and inter-individual multidimensionality. In H. Gulliksen & S. Messick (Eds.), *Psychometric scaling: Theory and applications* (pp. 155–167). New York: Wiley.

BIPOLAR SCALE

Bipolar scales are scales with the opposite ends represented by contrasting concepts such as *cold* and *hot*,

slow and *fast*, *antisocial* and *social*, or *introvert* and *extrovert*. There can be 2 to 13 points requested for such a scale.

—Tim Futing Liao

BISERIAL CORRELATION

Biserial correlation (r_{bis}) is a correlational index that estimates the strength of a relationship between an artificially dichotomous variable (X) and a true continuous variable (Y). Both variables are assumed to be normally distributed in their underlying populations. It is used extensively in item analysis to correlate a dichotomously scored item with the total score. Assuming that an artificially dichotomous item is scored 1 for a correct answer and 0 for an incorrect answer, the biserial correlation is computed from

$$r_{\text{bis}} = \frac{\bar{Y}_1 - \bar{Y}_0}{S_Y} \left(\frac{n_1 n_0}{un^2} \right),$$

where $\bar{Y}_1$ is the mean of Y for those who scored 1 on X , $\bar{Y}_0$ is the mean of Y for those who scored 0 on X ; S_Y is the standard deviation of all Y scores; n_1 is the number of observations scored 1 on X ; n_0 is the number of observations scored 0 on X ; $n = n_1 + n_0$, or the total number of observations; and u is the ordinate (i.e., vertical height or the probability density) of the standard normal distribution at the cumulative probability of $p_1 = n_1/n$. (i.e., the proportion of observations scored 1 on X).

Assuming that the following data set is obtained from 30 examinees who took the Science Aptitude Test,

their total test scores and item scores on Item #13 are as shown in the following table.

Treating the total score as the Y variable and the Item #13 score as the X variable, one calculates $\bar{Y}_0 = 12.93$, $\bar{Y}_1 = 29$, $S_Y = 11.92$, $n_0 = 15$, $n_1 = 15$, $n = 15 + 15 = 30$, and $p_1 = 15/30 = 0.50$. Taking p_1 to the standard normal distribution, one determines the ordinate, u , to be 0.3989. Substituting these values into the formula above, one obtains the biserial correlation as follows:

$$\begin{aligned} r_{\text{bis}} &= \frac{29.00 - 12.93}{11.92} \left[\frac{(15)(15)}{(0.3989)(30)^2} \right] \\ &= (1.3482)(0.6267) = 0.8449 \approx 0.85. \end{aligned}$$

The above value for biserial correlation can be interpreted as a measure of the degree to which the total score (continuous variable) differentiates between correct and incorrect answers on Item #13 (artificially dichotomous variable). The value of 0.85 suggests that the total score correlates strongly with the performance on this item. If the 15 students who failed this item had scored the lowest on the total test, the biserial correlation would have been 1.00 (the maximum). Conversely, if “pass” and “fail” on Item #13 were matched randomly with total scores, the biserial correlation would be 0.00 (the absolute minimum).

The biserial correlation is an estimate of the product-moment correlation, namely, Pearson’s r , if the normality assumption holds for X and Y population distributions. Its theoretical range is from -1 to $+1$. In social sciences research, r_{bis} computed from empirical data may be outside the theoretical range if the normality assumption is violated or when n is smaller than 15. Thus, r_{bis} may be interpreted as an estimate of Pearson’s r only if (a) the normality assumption is not

Examinee	Item #13 score (X)	Total score (Y)	Examinee	Item #13 score (X)	Total score (Y)	Examinee	Item #13 score (X)	Total score (Y)
1	0	8	11	1	14	21	0	28
2	0	12	12	1	13	22	1	33
3	0	6	13	0	10	23	1	32
4	0	12	14	0	9	24	1	32
5	0	8	15	0	8	25	1	33
6	0	8	16	1	33	26	0	34
7	0	8	17	0	28	27	1	35
8	0	11	18	1	29	28	1	34
9	1	13	19	1	30	29	1	38
10	0	4	20	1	29	30	1	37

violated, (b) the sample size is large (at least 100), and (c) the value of p_1 is not markedly different from 0.5. Biserial correlation should not be used in secondary analyses such as regression.

—Chao-Ying Joanne Peng

See also CORRELATION

REFERENCE

Glass, G. V. & Hopkins, K. D. (1984). *Statistical methods in education and psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

BIVARIATE ANALYSIS

Bivariate analysis is a statistical analysis of a pair of variables. Such analysis can take the form of, for example, CROSS-TABULATION, SCATTERPLOT, CORRELATION, ONE-WAY ANOVA, or simple REGRESSION. It is a step often taken after a UNIVARIATE ANALYSIS of the DATA but prior to a multivariate analysis.

—Tim Futing Liao

BIVARIATE REGRESSION. See
SIMPLE CORRELATION (REGRESSION)

BLOCK DESIGN

Block design is an experimental or QUASI-EXPERIMENTAL design in which units of study are grouped into categories or “blocks” in order to control undesired sources of variation in observed measurements or scores. An undesired source of variation is caused by nuisance variables that affect observed measurements or scores of observations but are not the main

interest or focus of a researcher. Nuisance variables are also called concomitant variables.

For example, in studying the effect of different treatments on depression, a researcher may wish to control the severity of a patient’s depression. Thus, in this study, the severity of depression is the nuisance variable. Levels of severity of depression are “blocks.” Let’s further assume that there are three treatments for depression and five levels of the severity of depression. The design layout looks as in the table below.

The statistical model assumed for a block design is

$$Y_{ij} = \mu + \alpha_i + \pi_j + \varepsilon_{ij},$$

$$(i = 1, \dots, n; \text{ and } j = 1, \dots, p), \text{ where}$$

Y_{ij} is the score of the i th block and the j th treatment condition

μ is the grand mean, therefore, a constant, of the population of observations

α_j is the treatment effect for the j th treatment condition; algebraically, it equals the deviation of the population mean (μ_j) from the grand mean (μ or $\mu_{..}$). It is a constant for all observations’ dependent score in the j th condition, subject to the restriction that all α_j sum to zero across all treatment conditions as in a fixed-effects design, or the sum of all α_j s equals zero in the populations of treatments, as in the random effects model

π_i is the block effect for population i and is equal to the deviation of the i th population mean (μ_i) from the grand mean ($\mu_{..}$). It is a constant for all observations’ dependent score in the i th block, normally distributed in the underlying population

ε_{ij} is the random error effect associated with the observation in the i th block and j th treatment condition. It is a random variable that is normally distributed in the underlying population of the i th block and j th treatment condition and is independent of π_i .

	Treatment 1	Treatment 2	Treatment 3	Row mean
Block 1 = most severe	Y_{11} (score)	Y_{12}	Y_{13}	$\bar{Y}_{1.}$
Block 2	Y_{21}	Y_{22}	Y_{23}	$\bar{Y}_{2.}$
Block 3	Y_{31}	Y_{32}	Y_{33}	$\bar{Y}_{3.}$
Block 4	Y_{41}	Y_{42}	Y_{43}	$\bar{Y}_{4.}$
Block 5 = least severe	Y_{51}	Y_{52}	Y_{53}	$\bar{Y}_{5.}$
Column mean	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{.3}$	Grand mean = $\bar{Y}_{..}$

A block design is also called *randomized block design* if observations are first randomly assigned to treatment conditions within each block. The procedure by which observations are assigned to treatment conditions is called “blocking.” There are various methods by which blocks are formed or observations are blocked on the nuisance variable. These methods attempt to achieve the goal of (a) including equal segments of the range of the nuisance variable in each block, or (b) including equal proportion of the observation population in each block. Most methods try to accomplish the second goal. Using the example above, a researcher may first rank order patients on their level of depression (such as the score on a depression scale). Second, he or she randomly assigns the top three scorers (most severely depressed patients) to the three treatment conditions within the first block. The second block is formed by the second set of three depressed patients who were judged to be less depressed than the top three but more depressed than the rest of sample. The process continues until all blocks are formed.

If more than one observation is assigned to the combination of block and treatment condition, the design is referred to as the *generalized block design*. Alternatively, the same patient (i.e., an observation) can serve as a block and is subject to all treatments in a random sequence. Some authors prefer to call this specific variation of block designs REPEATED-MEASURES DESIGN.

A randomized block design is appropriate for experiments that meet the usual ANOVA assumptions (a correct model, normality, equal variance, and independent random error) and the following additional assumptions:

- a. One treatment factor with at least two levels and a nuisance variable, also with two or more levels.
- b. The variability among units of analysis within each block should be less than that among units in different blocks.
- c. The observations in each block should be randomly assigned to all treatment levels. If each block consists of only one experimental unit that is observed under all treatment levels, the order of presentation of treatment levels should be randomized independently for each experimental unit, if the nature of treatment permits.

In general, there are three approaches to controlling the undesired source of variation in observed measurements:

- First, hold the nuisance variable constant; for example, use only patients who are mildly depressed.
- Second, assign observations randomly to treatment conditions with the assumption that known and unsuspected sources of variation among the units are distributed randomly over the entire experiment and thus do not affect just one or a limited number of treatment conditions in a specific direction. In other words, the principle of randomization will “take care of” the undesired source of variation in replications of the same study in the long run.
- Third, include the nuisance variable as one of the factors in the experiment, such as in block designs or analysis of covariance (ANCOVA). Thus, the main objective of block designs is to better investigate the main effect of the treatment variable (i.e., α_j), not the interaction between the treatment variable and a nuisance (also blocking) variable.

—Chao-Ying Joanne Peng

See also BLOCK DESIGN, EXPERIMENT, NUISANCE PARAMETERS

REFERENCES

- Huck, S. (2000). *Reading statistics and research* (3rd ed.). New York: Addison, Wesley, Longman.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.
- Kirk, R. E. (1999). *Statistics: An introduction* (4th ed.). Orlando, FL: Harcourt Brace.
- Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

BLUE. See BEST LINEAR UNBIASED ESTIMATOR

BMDP

BMDP is a software package containing 40+ powerful programs for statistical analysis, ranging from simple data description to complex multivariate analysis and time-series analysis. The output

produced by BMDP is very descriptive and largely self-explanatory.

—Tim Futing Liao

BONFERRONI TECHNIQUE

The Bonferroni technique is used to hold a familywise TYPE I ERROR rate to a preselected value when a family of related SIGNIFICANCE TESTS (or confidence intervals) is conducted. This situation, sometimes called MULTIPLE COMPARISONS, or, more generally, “simultaneous inference,” may arise when statistical tests are carried out on data from the same units (e.g., when people respond to several questionnaire items or when correlations among several variables are calculated on the same people), or when statistical estimates such as means or proportions are used more than once in a set of hypotheses (e.g., when a set of planned or post hoc comparisons are not statistically independent). Although the Bonferroni technique could be applied when the tests are independent, its use in such cases is extremely rare.

The Bonferroni technique has several advantages that make it an attractive alternative to other simultaneous inference procedures. Among these are that it may be applied to any statistical test resulting in a *P* VALUE, that it is very easy to use, and that it does not result in much wasted power when compared with other simultaneous significance testing procedures.

DEVELOPMENT

The basis of the Bonferroni technique is Bonferroni's inequality: $\alpha_f \leq \alpha_1 + \alpha_2 + \cdots + \alpha_m$. Because it is α_f that the researcher wishes to control, then choosing values for the several α_j terms that make $\alpha_1 + \alpha_2 + \cdots + \alpha_m = \alpha_f$ will result in control over α_f , the familywise Type I error rate. Usually, the several α_j values are set equal to each other, and each test is run at $\alpha_t = \alpha_f/m$. The value of α_f is therefore, at most, $m\alpha_t$.

APPLICATION

In order to apply the Bonferroni technique, we must have a family (i.e., set) of m (at least two) significance tests and will define two significance (α) levels: the per-test Type I error rate and the familywise Type I error rate. The per-test Type I error rate, α_t , is the

significance level of each statistical test (i.e., the null hypothesis is rejected if the p value of the test falls below α_t). The familywise Type I error rate, α_f , is the probability that one or more of the m tests in the family is declared significant as a Type I error (i.e., when its null hypothesis is true).

Use of the Bonferroni technique involves adjusting the per-test Type I error rate such that $\alpha_t = \alpha_f/m$. In practice, the familywise Type I error rate is usually held to .05. Thus, α_t is usually set at $.05/m$.

Although tables of these unusual percentiles of various distributions (e.g., t , chi-square) exist, in practice, the Bonferroni technique is most easily applied using the p values that are standard output in statistical packages. The p value of each statistical test is compared with the calculated α_t and if $p \leq \alpha_t$, the result is declared statistically significant; otherwise, it is not.

EXAMPLE

A questionnaire with four items has been distributed to a group of participants who have also been identified by gender. The intent is to test, for each item, whether there is a difference between the distributions of males and females. CHI-SQUARE TESTS based on the four two-way tables (gender by item response) will be used. However, if each test were conducted at the $\alpha = .05$ level, the probability of at least one of them reaching statistical significance by chance (assuming all null hypotheses true) is greater than .05 because there are four opportunities for a Type I error, not just one. In this example, there are $m = 4$ significance tests, so instead of comparing the p value for each test with .05, it is compared with $.05/4 = .0125$.

Similarly, if each of four experimental group means is to be compared with that of a control group, the Bonferroni technique could be used. Each of the four tests (t or F) would be accomplished by comparing the p value of the resulting statistic with .0125. Although the tests are correlated because they share a mean (that of the control groups), the familywise Type I error rate would be controlled to at most .05.

—William D. Schafer

BOOTSTRAPPING

Bootstrapping is a computer intensive, nonparametric approach to STATISTICAL INFERENCE. Rather

than making assumptions about the SAMPLING DISTRIBUTION of a statistic, bootstrapping uses the variability within a sample to estimate that sampling distribution empirically. This is done by randomly resampling with replacement from the sample many, many times in a way that mimics the original sampling scheme. There are various approaches to constructing CONFIDENCE INTERVALS with this estimated sampling distribution to make statistical inferences.

The goal of inferential statistical analysis is to make PROBABILITY statements about a POPULATION parameter, θ , from a STATISTIC, $\hat{\theta}$, calculated from sample data drawn from a population. At the heart of such an inference is the statistic's SAMPLING DISTRIBUTION, which is the range of values it could take on in a random sample from a given population and the probabilities associated with those values. In the standard parametric inferential statistics that social scientists learn in graduate school (with the ubiquitous *T*-TESTS and *P* VALUES), we get information about a statistic's sampling distribution from mathematical analysis. For example, the CENTRAL LIMIT THEOREM gives us good reason to believe that the sampling distribution of a sample mean is normal, with an expected value of the population mean and a STANDARD DEVIATION of approximately the standard deviation of the variable in the population divided by the square root of the sample size. However, there are situations where either no such parametric statistical theory exists or the assumptions needed to apply it do not hold. In these cases, one may be able to use bootstrapping to make a probability-based inference to the population parameter.

Bootstrapping is a general approach to statistical inference that can be applied to virtually any statistic. The basic procedure has two steps: (a) estimating the statistic's sampling distribution through resampling, and (b) using this estimated sampling distribution to construct confidence intervals to make inferences to population parameters.

First, a statistic's sampling distribution is estimated by treating the sample as the population and conducting a form of MONTE CARLO SIMULATION on it. This is done by randomly resampling with replacement a large number of samples of size n from the original sample of size n . Replacement sampling causes the resamples to be similar to, but slightly different from, the original sample because an individual case in the original sample may appear once, more than once, or not at all in any given resample.

Resampling should be conducted to mimic the sampling process that generated the original sample. Any stratification, WEIGHTING, clustering, stages, and so forth used to draw the original sample need to be used to draw each resample. In this way, the random variation that was infused into the original sample will be infused into the resamples in a similar fashion. The ability to make inferences from complex RANDOM SAMPLES is one of the central advantages of bootstrapping over parametric inference. In addition to mimicking the original sampling procedure, resampling ought to be conducted only on the random component of a statistical model. For example, this may be the raw data for simple descriptive statistics and survey samples or the ERROR term for a REGRESSION model.

For each resample, one calculates the sample statistic to be used in inference, $\hat{\theta}^*$. Because each resample is slightly and randomly different from each other resample, these $\hat{\theta}^*$ s will also be slightly and randomly different from one another. The central assertion of bootstrapping is that a relative FREQUENCY DISTRIBUTION of these $\hat{\theta}^*$ s is an estimate of the sampling distribution of $\hat{\theta}$, given the sampling procedure used to derive the original sample being mimicked in the resampling procedure.

To illustrate the effect of resampling, consider the simple example in Table 1. The original sample was drawn as a simple random sample from a standard normal distribution. The estimated mean and standard deviation vary somewhat from the population parameters (0 and 1, respectively) because this is a random sample. Note several things about the three resamples. First, there are no values in these resamples that do not appear in the original sample, because these resamples were generated from the original sample. Second, due to resampling with replacement, not every value in the original sample is found in each resample, and some of the original sample values are found in a resample more than once. Third, the sample statistics estimated from the resamples (in this case, the means and standard deviations) are close to, but slightly different from, those of the original sample. The relative frequency distribution of these means (or standard deviations or any other statistic calculated from these resamples) is the bootstrap estimate of the sampling distribution of the population parameter.

How many of these resamples and $\hat{\theta}^*$ s are needed to generate a workable estimate of the sampling distribution of $\hat{\theta}$? The estimate is asymptotically unbiased, but

Table 1 Example of Bootstrap Resamples

<i>Case Number</i>	<i>Original Sample [N(0,1)]</i>	<i>Resample #1</i>	<i>Resample #2</i>	<i>Resample #3</i>
1	0.697	-0.270	-1.768	-0.270
2	-1.395	0.697	-0.152	-0.152
3	1.408	-1.768	-0.270	-1.779
4	0.875	0.697	-0.133	2.204
5	-2.039	-0.133	-1.395	0.875
6	-0.727	0.587	0.587	-0.914
7	-0.366	-0.016	-1.234	-1.779
8	2.204	0.179	-0.152	-2.039
9	0.179	0.714	-1.395	2.204
10	0.261	0.714	1.099	-0.366
11	1.099	-0.097	-1.121	0.875
12	-0.787	-2.039	-0.787	-0.457
13	-0.097	-1.768	-0.016	-1.121
14	-1.779	-0.101	0.739	-0.016
15	-0.152	1.099	-1.395	-0.27
16	-1.768	-0.727	-1.415	-0.914
17	-0.016	-1.121	-0.097	-0.860
18	0.587	-0.097	-0.101	-0.914
19	-0.270	2.204	-1.779	-0.457
20	-0.101	0.875	-1.121	0.697
21	-1.415	-0.016	-0.101	0.179
22	-0.860	-0.727	-0.914	-0.366
23	-1.234	1.408	-2.039	0.875
24	-0.457	2.204	-0.366	-1.395
25	-0.133	-1.779	2.204	-1.234
26	-1.583	-1.415	-0.016	-1.121
27	-0.914	-0.860	-0.457	1.408
28	-1.121	-0.860	2.204	0.261
29	0.739	-1.121	-0.133	-1.583
30	0.714	-0.101	0.697	-2.039
<i>Mean</i>	-0.282	-0.121	-0.361	-0.349
<i>Standard deviation</i>	1.039	1.120	1.062	1.147

under what conditions is the variance small enough to yield inferences that are precise enough to be practical? There are two components to this answer. First, the asymptotics of the proof of unbiasedness for the bootstrap estimate of the sampling distribution appears to require an original sample size of 30 to 50, in terms of degrees of freedom. Second, the number of resamples needed to flesh out the estimated sampling distribution appears to be only about 1,000. With high-powered personal computers, such resampling and calculation requires a trivial amount of time and effort, given the ability to write an appropriate looping ALGORITHM.

After one estimates the sampling distribution of $\hat{\theta}$ with this resampling technique, the next step in bootstrap statistical inference is to construct confidence intervals using this estimate. There are several ways

to do this, and there has been some controversy as to which confidence interval approach is the most practical and statistically correct. Indeed, much of the discussion of the bootstrap in the 1980s and 1990s was devoted to developing and testing these approaches, which are too complicated to discuss in this article. The references listed below give details and instructions on these confidence interval approaches.

DEVELOPMENT

Bootstrapping is of a fairly recent vintage, with its genesis in a 1979 article by Bradley Efron in the *Annals of Statistics*, titled, "Bootstrap Methods: Another Look at the Jackknife." But as this title suggests, although Efron is widely regarded as the father of the bootstrap, the idea of assessing intrasample variability to

make nonparametric statistical inferences is not new. Statisticians have long understood the limitations of parametric inference and worked to overcome them. Although extremely powerful when appropriate, parametric statistical inference can be applied confidently to only a fairly narrow range of statistics in very specific circumstances. Ronald Fisher's exact test is one old way to make inferences without parametric assumptions, but it is limited in that it requires a complete delineation of all permutations of a given population. This is clearly not practical in most modern social science situations. Several techniques were discussed in the 1950s and 1960s that used the variability of a statistic in various subsamples to assess overall sampling variability. The most general of these approaches was jackknifing, which systematically eliminated cases from a sample one at a time and assessed the impact of this casewise elimination on a statistic's variability. But it was not until the advent of high-powered personal computers that Efron could take the leap to using unlimited numbers of randomly developed subsamples to estimate a statistic's sampling distribution, and thus the bootstrap was born.

APPLICATION

The two situations where bootstrapping is most likely to be useful to social scientists are (a) when making inferences using a statistic that has no strong parametric theory associated with it; and (b) when making inferences using a statistic that has strong parametric theory under certain conditions, but these conditions do not hold. For example, the bootstrap might be used in the former case to make inferences from indirect effects of path models, EIGENVALUES, the switch point in a switching regression, or the difference between two medians. This use of the bootstrap may become especially important as greater computer power and statistical sophistication among social scientists lead to the development of estimators designed to test specific hypotheses very precisely, but without any clear understanding of the properties of their sampling distributions. The second use of the bootstrap may be important as a check on the robustness of parametric statistical tests in the face of assumption violations. For example, because the asymptotic parametric theory for MAXIMUM LIKELIHOOD ESTIMATION (MLEs) appears to hold reasonably well only as sample sizes approach 100, the standard *t*-tests on MLEs may be inaccurate with a sample of, say, 75 cases. Bootstrap statistical

inference can be used in such a situation to see if this potential inaccuracy is a problem.

—Christopher Z. Mooney

REFERENCES

- Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide*. New York: Wiley-Interscience.
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1–26.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Thousand Oaks, CA: Sage.

BOUNDED RATIONALITY

Bounded rationality is a term used to describe a decision-making process. It is less a unified theory, and more a recognition that decision making in practice often does not conform to the concept of rationality that is the foundation of many formal models of behavior. The term was introduced by Herbert Simon in the 1950s (e.g., Simon, 1957) and is an active area of research today (e.g., Rubenstein, 1998). It is important to note that bounded rationality is not irrationality. Under bounded rationality, actors are still trying to make good decisions, but their decision-making process does not meet all of the criteria of full rationality.

Bounded rationality is perhaps best understood by comparison to the fully rational actors assumed in most formal models of behavior. Fully rational actors always make the best decision possible given the information available. Fully rational actors have clearly defined preferences, use all relevant information in an unbiased way, and are able to solve difficult optimization problems without error. Although most researchers who assume full rationality are aware that this is not a completely accurate description of decision making in reality, they often assume actors behave “as if” they were fully rational.

However, those who study bounded rationality have found many examples where actors do not behave “as if” they were fully rational. In reality, actors may not be capable of making the complex computations

required for fully rational behavior. For example, in an infinitely repeated game, full rationality requires that actors look an infinite number of periods into the future to plan their actions today, yet this most likely exceeds the computational abilities of even the most capable organizations or individuals. Under bounded rationality, actors still try to plan for the best actions today but are not capable of the complex computations undertaken by fully rational actors. Perhaps the most common distinction between bounded and unbounded rationality is due to Simon, who holds that under full rationality, actors optimize, finding the best solution to a problem, whereas under bounded rationality, actors often satisfice, adopting a solution that is “close enough.”

Other examples of boundedly rational behavior have been documented, such as a tendency to simplify problems and framing effects. Fully rational actors are indifferent to logically equivalent descriptions of alternatives and choice sets. However, in a well-known EXPERIMENT, Tversky and Kahneman (1986) demonstrated that the framing of a decision problem can have an impact on the choices individuals make. Subjects in the experiment were told that an outbreak of a disease will cause 600 people to die in the United States. The subjects were divided into two groups. The first group of subjects was asked to choose between two mutually exclusive treatment programs (A and B) that would yield the following results:

- A. Two hundred people will be saved.
- B. Six hundred people will be saved with probability 1/3; nobody will be saved with probability 2/3.

The second group of subjects was asked to choose between two mutually exclusive programs (C and D) with the following results:

- C. Four hundred people will die.
- D. Nobody will die with probability 1/3; all six hundred will die with probability 2/3.

It is clear that A is equivalent to C, and B is equivalent to D. However, 72% of individuals in the first group chose A, while 78% of individuals in the second group chose D. This result is more evidence that there are differences between actual behavior and the assumptions of full rationality, and thus, many

actors in decision-making problems are best described as boundedly rational.

—Garrett Glasgow

REFERENCES

- Rubenstein, A. (1998). *Modeling bounded rationality*. Cambridge: MIT Press.
- Simon, H. A. (1957). *Models of man*. New York: Wiley.
- Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, 59, 251–278.

BOX-AND-WHISKER PLOT.

See BOXPLOT

BOX-JENKINS MODELING

The methodology associated with the names of Box and Jenkins applies to TIME-SERIES data, where observations occur at equally spaced intervals (Box & Jenkins, 1976). Unlike DETERMINISTIC MODELS, Box-Jenkins models treat a time series as the realization of a STOCHASTIC process, specifically as an ARIMA (autoregressive, integrated, moving average) process. The main applications of Box-Jenkins models are forecasting future observations of a time series, determining the effect of an intervention in an ongoing time series (INTERVENTION ANALYSIS), and estimating dynamic input-output relationship (transfer function analysis).

ARIMA MODELS

The simplest version of an ARIMA model assumes that observations of a time series (z_t) are generated by random shocks (u_t) that carry over only to the next time point, according to the moving-average parameter (θ). This is a first-order moving-average model, MA(1), which has the form (ignore the negative sign of the parameter)

$$z_t = u_t - \theta u_{t-1}. \quad (1)$$

This process is only one short step away from pure randomness, called “white noise.” The memory today (t) extends only to yesterday’s news ($t - 1$), not to any before then. Even though a moving-average model can be easily extended, if necessary, to accommodate a

larger set of prior random shocks, say u_{t-2} or u_{t-3} , there soon comes the point where the MA framework proves too cumbersome. A more practical way of handling the accumulated, though discounted, shocks of the past is by way of the AUTOREGRESSIVE model. That takes us to the “AR” part of ARIMA:

$$z_t = \phi z_{t-1} + u_t. \quad (2)$$

In this process, AR(1), the accumulated past (z_{t-1}), not just the most recent random shock (u_{t-1}), carries over from one period to the next, according to the autoregressive parameter (ϕ). But not all of it does. The autoregressive process requires some leakage in the transition from time $t - 1$ to time t . As long as ϕ stays below 1.0, the AR(1) process meets the requirements of stationarity (constant mean level, constant variance, and constant covariance between observations).

Many time series in real life, of course, are not stationary. They exhibit trends (long-term growth or decline) or wander freely, like a RANDOM WALK. The observations of such time series must first be transformed into stationary ones before autoregressive and/or moving-average components can be estimated. The standard procedure is to difference the time series z_t ,

$$\nabla z_t = z_t - z_{t-1}, \quad (3)$$

and to determine whether the resulting series (∇z_t) meets the requirements of stationarity. If so, the original series (z_t) is considered integrated at order 1. That is what the “I” in ARIMA refers to. I simply counts the number of times a time series must be differenced to achieve stationarity. One difference ($I = 1$) is sufficient for most nonstationary series.

The analysis of a stationary time series then proceeds from model identification, through parameter estimation, to diagnostic checking. To identify the dynamic of a stationary series as either autoregressive (AR) or moving average (MA), or as a combination of both—plus the order of those components—one examines the AUTOCORRELATIONS and partial autocorrelations of the time series up to a sufficient number of lags. Any ARMA process leaves its characteristic signature in those correlation functions (ACF and PACF). The parameters of identified models are then estimated through iterative MAXIMUM LIKELIHOOD ESTIMATION procedures. Whether or not this model is adequate for capturing the ARMA dynamic of the time series depends on the error diagnostic. The Ljung-Box

Q-statistic is a widely used summary test of white-noise residuals.

FORECASTING

Box-Jenkins models were designed to avoid the pitfalls of forecasting with deterministic functions (extrapolating time trends). An ARIMA forecast is derived from the stochastic dynamic moving a time series. Once the parameters of the model have been obtained, no other data are required to issue the forecast. Forecasts come automatically with the model estimates, and they arrive before the event to be predicted takes place. Hence, ARIMA forecasts are true ex-ante, unconditional forecasts that incorporate all the information about the phenomenon’s dynamic behavior. For example, an AR(2) model of the U.S. presidential vote, estimated with aggregate results from 1860 to 1992, was able to forecast quite early the Democratic victory in 1996 (Norpoth, 1995).

INTERVENTION EFFECTS

There are many forms of behavior and performance that government intends to change by adopting policies. Wherever those outcomes are measured frequently, Box-Jenkins models are able to deliver statistical estimates of the effect of such interventions (Box & Tiao, 1975; Clarke, Norpoth, & Whiteley, 1998). The basic form of an intervention model, say for a policy aimed at reducing traffic fatalities, is as follows:

$$y_t = [\omega/(1 - \delta B)] I_t + N_t, \quad (4)$$

where

y_t = fatalities in the t th month,

I_t = the intervention variable (1 for month of policy adoption, and 0 elsewhere),

ω = the step change associated with the intervention variable,

δ = rate of decay of the step change,

B = backward-shift operator (such that, for example, $BI_t = I_{t-1}$),

N_t = “noise” modeled as an ARIMA process.

This model does not require the impact of the intervention to be durable, nor that the impact is gone the moment the intervention is over. By modeling the time series as an ARIMA process, one avoids

confounding the alleged intervention impact with many other possible effects. Such a randomization precludes mistaking a trend for an intervention effect.

TRANSFER FUNCTIONS

Instead of a discrete intervention changing the course of a time series, consider another time series continuously affecting the ups and downs of the time series of interest. This influence need not be instant, but may take time to materialize fully. This can be expressed with a transfer function model (Box & Jenkins, 1976):

$$y_t = [\omega/(1 - \delta B)] x_t + e_t. \quad (5)$$

In this model, ω and δ describe an influence constantly at work between X and Y . The moment X moves, Y does, too, with a magnitude set by the ω parameter. Then, a moment later, that movement of X triggers yet a further reaction in Y , according to the δ parameter, and so on. To identify a dynamic model of this sort for variables that are each highly autocorrelated, Box and Jenkins (1976) introduced the following strategy: “Prewhiten” the X series by removing its ARIMA dynamic; then, filter the Y series accordingly; and then obtain cross-correlations between the transformed time series. This is a drastic strategy that may end up “throwing the baby out with the bath water.” Caution is advised.

Although Box-Jenkins models cover a wide array of time-series problems, this is not a methodology where something always shows up. The user faces the risk of coming up empty-handed. Time series is a domain where near-perfect correlations can dissolve into randomness in a split second. Difference a random walk, where successive observations are perfectly correlated, and you get white noise, where the correlation is zero.

—Helmut Norpoth

REFERENCES

- Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.
- Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70, 70–79.
- Clarke, H. D., Norpoth, H., & Whiteley, P. F. (1998). It's about time: Modeling political and social dynamics. In E. Scarborough & E. Tanenbaum (Eds.), *Research strategies*

in the social sciences (pp. 127–155). Oxford, UK: Oxford University Press.

Granger, C. W. J., & Newbold, P. (1986). *Forecasting economic time series* (2nd ed.). New York: Academic Press.

Mills, T. C. (1990). *Time series techniques for economists*. Cambridge, UK: Cambridge University Press.

Norpoth, H. (1995). Is Clinton doomed? An early forecast for 1996. *Political Science & Politics*, 27, 201–206.

BOXPLOT

Boxplots were developed by Tukey (1977) as a way of summarizing the midpoint and range of a distribution and comparing it with other distributions, or comparing the distributions of a given variable grouped by another, categorical, variable. The size of the box indicates the range of the middle 50% of values of the chosen variable, that is, the range from the 25th to the 75th quartile, or, as it is also known, the INTERQUARTILE RANGE. A line across the box represents the MEDIAN value. *Whiskers* extend from the box to reach the furthest values at either end, excluding outliers. That is, they reach to the furthest values that fall from either end of the box within 1.5 times the size of the box. These limits to the length of the whiskers are known as *fences*. The furthest values reached by the whiskers will not necessarily reach to the fences. Whether they do or not will depend on the shape of the distribution and the amount of dispersion. OUTLIERS are considered to be values that fall beyond the fences and may be noted separately and labeled individually. *Extremes* are an extreme version of outliers that fall further than three times the size of the box from either end of the box. Again, they may be noted and labeled individually. Thus, boxplots provide, in condensed, graphical form, a large amount of key information about a distribution and are particularly useful for identifying its extremes and outliers.

Figure 1 provides an example of a boxplot comparing the earnings of men and women in Britain in 1999. The boxplot indicates the shape of the earnings distribution. We can see that it is skewed to the right (it clusters closer to 0) for both men and women. However, the median for women falls lower in the box, indicating a greater clustering of lower incomes. In addition, the whole distribution is situated lower for women, indicating that their earnings are systematically lower. There are a number of outliers (labeled by their case number), showing that a few people have

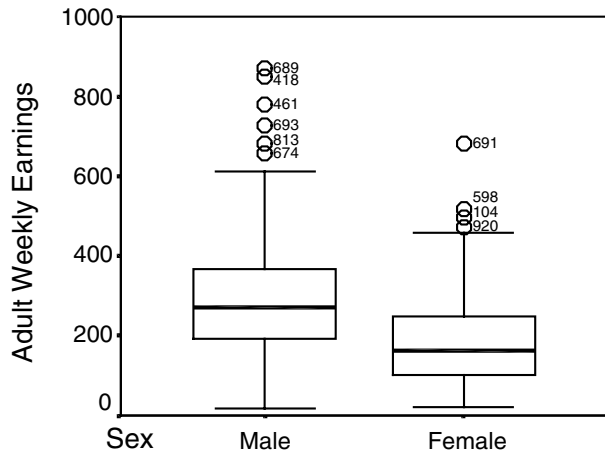


Figure 1 Boxplot of Male and Female Weekly Earnings

very high earnings. It is interesting to note, though, that most of the female outliers fall within the fence of the male distribution.

Figure 2 is an example of a clustered bar chart, which breaks down the information in Figure 1 by education level, but supplying all the same information as in Figure 1 for each subcategory. It explores whether the patterns illustrated in Figure 1 are replicated when observations are clustered by sex *within* education categories. It shows how compressed the earnings distribution is for women without a higher qualification and the absence of outliers for this group. It also shows how the interquartile range narrows steadily across the four groups, indicating

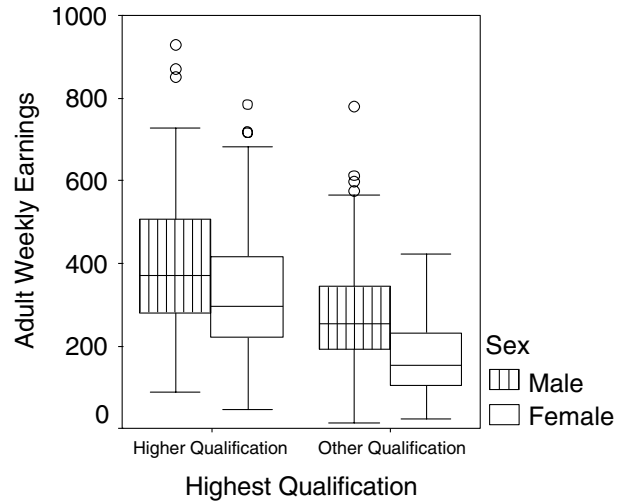


Figure 2 Boxplot of Adult Weekly Earnings by Sex and Highest Qualification, Britain 1999

most heterogeneity in earnings among highly qualified men and most homogeneity among less well-qualified women. It demonstrates the power of graphical representation not only to convey essential information about our data, but also to encourage us to think about it in an exploratory way.

—Lucinda Platt

REFERENCES

- Marsh, C. (1988). *Exploring data: An introduction to data analysis for social scientists*. Cambridge, UK: Polity.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

C

CAIC. See GOODNESS-OF-FIT MEASURES

CANONICAL CORRELATION ANALYSIS

The basic idea of canonical correlation analysis is clearly illustrated by Russett (1969), who analyzed the ASSOCIATION between some economic and political characteristics of 47 countries. Five indicators were used for measuring economic inequality: the division of farmland, the GINI COEFFICIENT, the percentage of tenant farmers, the gross national product (GNP) per capita, and the percentage of farmers. Russett measured political instability with four indicators: the instability of leadership, the level of internal group violence, the occurrence of internal war, and the stability of democracy. The research hypothesis was that Alexis de Tocqueville was right: There is not one nation that is capable of maintaining a democratic form of government for an extensive period of time if economic resources are unevenly distributed among its citizens. In other words, there is a significant association between the economic *inequality* and the political *instability* of nations.

The two theoretical concepts in this research problem are “economic inequality” (X^*) and “political instability” (Y^*). They are called canonical variables with an expected high CORRELATION, known as *canonical correlation*. The first canonical variable, X^* , is measured by $p = 5$ indicators, X_1 to X_5 , and we will

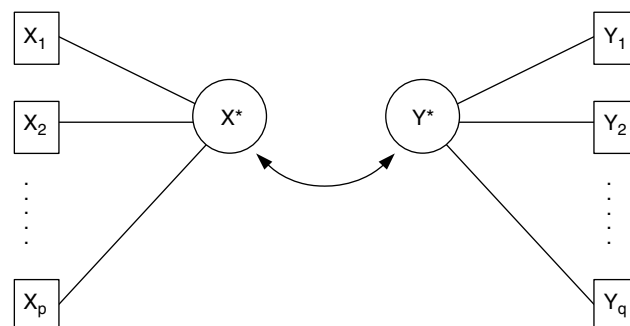


Figure 1 Canonical Correlation

consider X^* as a linear combination (a weighted sum) of these X variables. In an analogous fashion, Y^* , the second canonical variable, is a linear combination of the $q = 4$ indicators, Y_1 to Y_4 . In the most general case, in which the X set contains p variables and the Y set q variables, the diagram can be represented as in Figure 1.

In Russett’s (1969) research problem, the canonical correlation between the canonical variables X^* and Y^* is in fact a causal relationship because he asserts that nations with greater economic inequality display greater political instability. The arrow in the other direction, according to which the political characteristics would influence the economic characteristics, is not implied in de Tocqueville’s or Russett’s view. However, considering the fact that a canonical correlation is a “correlation,” the statistical analysis is not asymmetrical. For this reason, we do not draw a causal arrow from X^* to Y^* in the figure but a double-sided curved arrow, indicating that the question of CAUSALITY remains open.

THE MODEL OF CANONICAL CORRELATION ANALYSIS

In canonical correlation analysis, we want to see if there is a significant association between a set of X variables and a set of Y variables. For this reason, we look for a linear combination of the X set, $X^* = a_1X_1 + a_2X_2 + \cdots + a_pX_p$ and a linear combination of the Y set, $Y^* = b_1Y_1 + b_2Y_2 + \cdots + b_qY_q$ in such a way that X^* and Y^* are maximally correlated. The two linear combinations, X^* and Y^* , are not observed. In our example, they have been given a name a priori: economic inequality and political instability. Sometimes such an a priori theory is absent at the start of research, and a name must be devised afterwards.

To find these two linear combinations (canonical variables X^* and Y^*), the a and b weights must be calculated. Canonical correlation analysis aims at determining these weights in such a way that the canonical correlation ρ is as high as possible. The square of this canonical correlation is the proportion of variance in one set (e.g., political characteristics) that is explained by the variance in the other set (e.g., economic characteristics).

Just as in PRINCIPAL COMPONENTS ANALYSIS, in which not one but several uncorrelated components can be determined, here various pairs of canonical variables can be found. Each pair is uncorrelated with the preceding pair and is calculated each time so that the canonical correlation is maximal. The number of pairs of canonical variables or, which comes down to the same, the number of canonical correlations is equal to the number of variables in the smallest set, that is, $\min[p, q]$.

The calculation of each of the canonical correlations and of the a and b weights will be done by examining the eigenstructure of a typical matrix. In DISCRIMINANT ANALYSIS and principal components analysis, this typical matrix is $\mathbf{W}^{-1}\mathbf{B}$ and $\mathbf{R}$, respectively. In canonical correlation analysis this matrix is $\mathbf{R}_{yy}^{-1}\mathbf{R}'_{xy}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$. In this matrix notation, a distinction is made between different types of within- and between-correlation matrices. Between correlations are calculated ($\mathbf{R}'_{xy}$ and $\mathbf{R}_{xy}$), and a control for within correlations is built in ($\mathbf{R}_{yy}^{-1}$ and $\mathbf{R}_{xx}^{-1}$). We will now explain this in more detail.

THREE TYPES OF CORRELATIONS

A legitimate method for examining the correlations between the economic and political characteristics is

an analysis of the separate correlations between, for example, the Gini index and the (in)stability of leadership (X_1, Y_1) or between the percentage of farmers and the level of violence (X_2, Y_2). These are the correlations between X and Y variables and are therefore called *between correlations*.

Canonical correlation analysis, however, goes a step further than this examination of separate correlations.

First of all, in canonical correlation analysis, one aims to examine to what degree two sets— X variables on one hand and Y variables on the other—show association “as a whole.” To this end, each set is replaced by a linear combination. The correlation between these two linear combinations is the objective of the research.

Second, in canonical correlation analysis, one controls for the associations within each of the sets, as it is possible that the variables within the X set of economic characteristics are also correlated in pairs. These types of correlations are called *within correlations*. They form a problem that is comparable to multicollinearity in multiple regression analysis. There, the regression coefficients that are calculated are partial coefficients (i.e., we control each variable for association with the other variables in the X set). Something similar is done in canonical correlation analysis, only here we do not just control on the side of the X set but on the side of the Y set of political characteristics as well because we now have various Y variables that could also be correlated in pairs.

We can, therefore, distinguish three types of correlations:

1. Correlations between X variables; the correlation matrix is $\mathbf{R}_{xx}$.
2. Correlations between Y variables; the correlation matrix is $\mathbf{R}_{yy}$.
3. Correlations between X and Y variables; the correlation matrix is $\mathbf{R}_{xy} = \mathbf{R}'_{yx}$.

The fact that canonical correlation analysis aims to examine the between correlations and not the within correlations can be illustrated more clearly by comparing this technique to double principal components analysis (PCA twice). PCA looks for the correlations within one set of variables (i.e., within correlations). If we were to perform two PCAs, one on the X set and one on the Y set, we would obtain a first principal component in each PCA. The correlation between these two principal components is absolutely not the same as the (first) canonical correlation because the correlation between the two components measures to what

degree the within associations in the set of economic characteristics are comparable to the within associations in the set of political characteristics. Canonical correlation, on the other hand, measures the degree to which the economic and political characteristics are correlated, controlling for (i.e., assuming the absence of) within associations within the economic set, on one hand, and within the political set, on the other.

GEOMETRIC APPROACH

To explain canonical correlation from a geometric viewpoint, we create two coordinate systems: one with the x variables as axes and one with the y variables as axes. Canonical correlation analysis aims at finding a first pair of canonical variables, one X^* in the space of X variables and one Y^* in the space of Y variables, in such a way that the correlation between X^* and Y^* is maximal. Afterwards, one looks for a second pair of canonical variables, uncorrelated with the first pair, in such a way that the (canonical) correlation between this second pair is as high as possible. The high canonical correlations between the first pair, X_1^* and Y_1^* , as well as the second pair, X_2^* and Y_2^* , can be seen by looking at the coordinate systems in which the canonical variables are axes. The fact that canonical variables of different pairs—for example, X_1^* and X_2^* (and also Y_1^* and Y_2^*)—are uncorrelated does not necessarily mean that they are orthogonal in the space of x variables. It does mean that the regression line in (X_1^*, X_2^*) space has a zero slope.

OBJECTIVES OF THE TECHNIQUE

Three goals are pursued in canonical correlation analysis. With Russett's (1969) study in mind, these goals can be described as the following:

1. We look for a first pair of linear combinations, one based on the set of economic variables and the other based on the set of political variables, in such a way that the correlation between both of the linear combinations is maximal. Next we look for a second pair of linear combinations, also maximally correlated and uncorrelated with the first pair. We do this as many times as there are variables in the smallest set. We call the linear combinations *canonical variables*. The weights are known as *canonical weights*. The maximized correlations are called *canonical correlations*. On the basis of the canonical weights, we attempt to interpret the associations "between" the sets. One can

also attempt to give a name to each canonical variable per pair.

2. Making use of the canonical weights, the scores of the canonical variables are calculated. The correlations between the original variables and the canonical variables often offer additional possibilities for interpretation. These are called *structure correlations*.

3. We check to see whether the associations between the two sets can be generalized to the population of the countries. This is done by testing the canonical correlations for significance. We can perform this test for all of the canonical correlations together or for each of the separate canonical correlations or their subgroups. Because the first canonical correlation is the largest and the following correlations become smaller and smaller, it seems obvious to test these separately and only to interpret the ones that give us the most significant result.

A CAUSAL APPROACH TO CANONICAL CORRELATION ANALYSIS

In 1968, Stewart and Love developed an asymmetrical version of canonical correlation analysis. The canonical correlations calculated above are, after all, symmetrical coefficients. For example, the square of the first canonical correlation (i.e., the first eigenvalue λ_1), which represents the squared correlation between the first pair of canonical variables, can be interpreted as the proportion of variance in the Y set (political characteristics) that is explained by the X set (economic characteristics). One could, however, just as easily reason the other way round: λ_1 is also the proportion of variance in economic inequality that is explained by political instability. In other words, no distinction is made between independent and dependent variables, between cause and effect.

For this reason, Stewart and Love proposed an *index of redundancy*. This index measures the degree to which the Y set can be reconstructed on the basis of the knowledge of the X set. It appears that this index is equal to the arithmetic mean of all multiple determination coefficients that are obtained by repeated multiple regression analyses with one Y variable on the function of the entire X set.

THE SIGNIFICANCE TESTS

To test whether the canonical correlations are significant, we make use of Wilk's lambda, Λ , a

measure—just like Student's t , Hotelling's T , and Fisher's F —that confronts the ratio of the between dispersion and the within dispersion from the sample with this ratio under the null hypothesis. Wilk's Λ is the multivariate analog of this ratio in its most general form, so that t , T , and F are special cases of Λ . This measure is therefore used for more complex multivariate analysis techniques, such as multiple discriminant analysis, multivariate analysis of variance and covariance, and canonical correlation analysis.

In canonical correlation analysis, Wilk's Λ can be calculated as the product of error variances, in which the error variances are determined by the eigenvalues (i.e., the squared canonical correlations and, therefore, the proportions of explained variance) subtracted from 1. Bartlett (1947) constructed a V measure that is distributed as chi-square with pq degrees of freedom, where p and q represent the number of variables of the X set and the Y set:

$$V = -[n - 1 - (p + q + 1)/2] \ln \lambda.$$

When, for pq degrees of freedom, this value is greater than the critical chi-square value that is found under the null hypothesis for a significance level of $\alpha = 0.05$, then there is a significant association between the economic inequality and political instability of countries. The hypothesis of de Tocqueville and Russett is then empirically supported.

The V measure can also be split additively, so that a separate test can be performed as to whether the significant correlation between the economic and political characteristics also holds for the second and following canonical correlations.

EXTENSIONS

Cases can arise in applied multivariate research in which the research problem is composed of more than two sets of characteristics for the same units; for example, next to an economic X set and a political Y set, there is also a Z set of sociocultural characteristics. Generalized canonical correlation is the name given to procedures used to study correlation between three or more sets.

In the foregoing discussion, it was taken for granted that the variables were all measured at the interval or ratio level. It was also assumed that the relationships were all linear. These two requirements can, however, be relaxed. The computer program CANALS performs a canonical analysis, which is nonmetric as well as

nonlinear. This promising new development deserves our attention. The interested reader is referred to Gifi (1980).

—Jacques Tacq

REFERENCES

- Bartlett, M. S. (1947). Multivariate analysis. *Journal of the Royal Statistical Society, Series B*, 9, 176–197.
- Cooley, W., & Lohnes, P. (1971). *Multivariate data analysis*. New York: John Wiley.
- Gifi, A. (1980). *Nonlinear multivariate analysis*. Leiden, The Netherlands: Leiden University Press.
- Green, P. (1978). *Analyzing multivariate data*. Hinsdale, IL: Dryden.
- Russett, B. (1969). Inequality and instability: The relation of land tenure to politics. In D. Rowney & J. Graham (Eds.), *Quantitative history: Selected readings in the quantitative analysis of historical data* (pp. 356–367). Homewood, IL: Dorsey.
- Stewart, D. K., & Love, W. A. (1968). A general canonical correlation index. *Psychological Bulletin*, 70, 160–163.
- Van de Geer, J. (1971). *Introduction to multivariate analysis for the social sciences*. San Francisco: Freeman.

CAPI. See COMPUTER-ASSISTED DATA COLLECTION

CAPTURE-RECAPTURE

Capture-recapture, also called *mark-recapture*, is a method for estimating the size of a POPULATION as follows: Capture and mark some members of the population. Take a random SAMPLE from the population (recapture), and count the members of the sample with marks. The fraction marked in the sample estimates the fraction marked in the population. The estimated size of the population is the number marked in the population divided by the fraction marked in the sample. For example, suppose that 1,000 members of a population were marked and that in a random sample of 500 members of the population, 200 had marks. Then the 1,000 marked members of the population comprise about $200/500 = 0.4$ of the population, so the population size is about $1,000/0.4 = 2,500$. Laplace was the first to use capture-recapture to estimate the size of a population—that of France in 1786 (Cormack, 2001). Capture-recapture has several crucial assumptions:

1. The sample (recapture) is a simple random sample from the population. So every population member must have the same chance of being in the sample, and the chance that a given member is in the sample cannot depend on whether the member was marked. This is called the independence assumption; violations cause correlation bias.

2. The population of marked and unmarked individuals is constant between capture and recapture (no births, deaths, immigration, or emigration).

3. The marks are unambiguous.

Refinements of the method try to account for population changes, unequal PROBABILITIES of recapture, time-varying probabilities of recapture, "trap effects," and so on (Cormack, 2001; Otis, Burnham, White, & Anderson, 1978; Pollock, Nichols, Brownie, & Hines, 1990). Capture-recapture is sometimes used to estimate the size of animal populations and has been adapted to estimate other population parameters, such as survival rates (Lebreton, Burnham, Clobert, & Anderson, 1992). Capture-recapture is the basis for a method that has been proposed to adjust the U.S. census for undercount (see CENSUS ADJUSTMENT).

—Philip B. Stark

REFERENCES

- Cormack, R. (2001). Population size estimation and capture-recapture methods. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 17). Amsterdam: Elsevier.
- Lebreton, J.-D., Burnham, K. P., Clobert, J., & Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies. *Ecological Monographs*, 62, 67–118.
- Otis, D. L., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, 62, 1–135.
- Pollock, K. H., Nichols, J. D., Brownie, C., & Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs*, 107, 1–97.

CAQDAS (COMPUTER-ASSISTED QUALITATIVE DATA ANALYSIS SOFTWARE)

Computer-assisted qualitative data analysis software (CAQDAS) is a term introduced by Fielding and

Lee (1998) to refer to the wide range of software now available that supports a variety of analytic styles in QUALITATIVE RESEARCH. The software does not "do" the analysis; it merely assists. It needs people to read, understand, and interpret the data. An essential part of qualitative data analysis is the effective handling of data, and this requires careful and complex management of large amounts of texts, video, codes, memos, notes, and so on. Ordinary word processors and database programs as well as dedicated text management programs have been used for this to good effect. However, a core activity in much qualitative analysis is CODING: the linking of one or more passages of text (or sequences of video) in one or more different documents or recordings with named analytic ideas or concepts. A second generation of CAQDAS introduced facilities for coding and for manipulating, searching and reporting on the text, images, or video thus coded. Such code-and-retrieve functions are now at the heart of the most commonly used programs.

PROGRAM FUNCTIONS

Most programs work with textual data. Although it is not necessary to transcribe ethnographic notes, narratives, and so forth, it is usually advantageous because most software supports the rapid recontextualization of extracted and/or coded passages by displaying the original text. Most programs support the coding of text, and such codes can be arranged hierarchically, modified, and rearranged as the analysis proceeds. Some software (e.g., MaxQDA, Qualrus, ATLAS.TI, and NVIVO) can work with rich text that includes different fonts, styles, and colors. Several programs support the analysis of digitized audio, images, and video, in some cases allowing the direct coding of digitized material (e.g., HyperRESEARCH, Atlas.ti, CI-SAID, AnnoTape, and the Qualitative Media Analyser). So far, there is no software that can automatically transcribe from audio recordings. Even the best speech recognition software still relies on high-quality recordings and familiarization with one speaker and cannot cope with multiple respondents' voices.

Most software allows the writing of MEMOS, code definitions, FIELDNOTES, and other analytic texts as part of the process of analysis, and some (e.g., NVivo, NUD*IST/N6, Atlas.ti, and Qualrus) allow this text to be linked with data and codes in the

project database. A key feature of many programs is the ability to search online text for words and phrases. Typically, such lexical searching allows the marking of all the matching text wherever it is found in the documents. Some programs support the searching of the database using the coded text. The most sophisticated of these allow complex searches involving various Boolean and proximity combinations of words, phrases, coded text, and VARIABLES (e.g., NVivo, MaxQDA, WINMAX, Atlas.ti, NUD*IST/N6, HyperRESEARCH, and Qualrus).

A third generation of software has extended these functions in two ways. Some assist analytic procedures with a variety of facilities to examine features and relationships in the texts. Such programs (Qualrus, HyperRESEARCH, Ethno, and AQUAD Five) are often referred to as theory builders or model builders because they contain various tools that assist researchers to develop theoretical ideas and test a HYPOTHESIS (Kelle, 1995). Another example in some programs is a model-building facility that uses diagrams linked to the main database of codes, text, or images (e.g. Atlas.ti, Qualrus, and NVivo). Second, some programs enable the exchange of data and analyses between collaborating researchers. Some do this by the export and import of data and analyses in a common format such as XML (e.g., Tatoe and Atlas.ti). Others are able to merge project databases from different researchers (e.g., NVivo, Atlas.ti, and NUD*IST/N6).

CHOOSING SOFTWARE

As Weitzman and Miles (1995) have shown, CAQDAS programs differ in facilities and operation. Which program is appropriate will depend partly on practical issues such as funding and available expertise, but the type of analysis needed is also relevant. Some software programs are better for large data sets with structured content and for linking with quantitative data (NUD*IST/N6), but others are better at dealing with digitized video and audio (Qualitative Media Analyser, AnnoTape). Some allow complex, hierarchically and multiply coded text; others allow only simple coding of passages. Analytic style matters too. Styles involving thematic analysis such as GROUNDED THEORY and various kinds of PHENOMENOLOGY are well supported. However, case-based approaches and NARRATIVE, DISCOURSE, and CONVERSATION ANALYSIS are less well supported.

ADVANTAGES AND DISADVANTAGES

Qualitative data tend to be voluminous, and it is all too easy to produce partial and biased analyses. The use of CAQDAS helps, not least because it removes much of the sheer tedium of analysis. It is easier to be exhaustive, to check for NEGATIVE CASES, and to ensure that text has been coded in consistent and well-defined ways. An audit trail of analysis can be kept so that emerging analytic ideas can be recalled and justified. Most programs allow the production of simple statistics so that the researcher can check (and give evidence) whether examples are common or idiosyncratic. In general, the advantage of CAQDAS is not that it can do things, such as working in teams, with large data sets or with audio and video, that could not be done any other way but rather that, without the software, such things are impossibly time-consuming.

Nevertheless, the use of CAQDAS remains contested. Early users felt that they were closer to the words of their respondents or to their field notes when using paper-based analysis rather than computers. It is likely that this was because users were familiar with traditional approaches but new to computers. Now, most software has good facilities for recontextualization, and there is little reason for those using CAQDAS to feel distant from the data. Other critics suggest that CAQDAS reinforces certain biases such as the dominance of grounded theory and the code-and-retrieve approach. Although it is true that much software has been inspired by grounded theory and other approaches that give a primacy to coding text, the programs are now more sophisticated and flexible and have become less dedicated to any one analytic approach. Despite this, there is evidence that most published research describing the use of CAQDAS remains at a very simple level of pattern analysis and shows little evidence of good reliability or validity. As users become more experienced and more researchers use the software, this is likely to change, and innovative use of the software will increase.

—Graham R. Gibbs

REFERENCES

- Fielding, N. G., & Lee, R. M. (1998). *Computer analysis and qualitative research*. London: Sage.
- Kelle, U. (Ed.). (1995). *Computer-aided qualitative data analysis: Theory, methods and practice*. London: Sage.

Weitzman, E. A., & Miles, M. B. (1995). *Computer programs for qualitative data analysis: A software source book*. Thousand Oaks, CA: Sage.

CARRYOVER EFFECTS

Carryover effects occur when an experimental treatment continues to affect a participant long after the treatment is administered. Carryover effects are ubiquitous because virtually everything that happens to an organism continues to affect the organism in some fashion for some time, like ripples on the surface of a pond radiating from a thrown stone. Some treatments may produce carryover effects that are fairly modest in size and/or fairly limited in duration, but others may produce carryover effects that are substantial and/or sustained.

The extended impact of a prescription drug provides the most accessible example of carryover effects because people are well aware of the possibility of a drug continuing to exert its influence for a substantial period of time after it is administered. In fact, many drug manufacturers advertise claims about the extended effect of their drugs. People are also aware of the potential complications from the interaction of different drugs, leading them to avoid taking one drug while another is still present in their systems.

Researchers can eliminate or minimize the impact of carryover effects in their studies by means of different strategies. The most obvious approach would be to eliminate the impact of carryover effects by using an independent groups design. In such a design, only one treatment is administered to each participant, data are collected, and the participant's role in the study is complete. Thus, any carryover effects that occur would have no impact on the data collected in other conditions of the study (although the treatment may continue to affect the participant).

Carryover effects are of great concern in experiments using REPEATED-MEASURES DESIGNS because participants are exposed to a series of treatments, after each of which data are collected. Thus, carryover effects from an early treatment could continue to affect the participant's responses collected after a subsequent treatment. In such a case, of course, the response to the later treatment cannot be interpreted as unambiguously due to that treatment but is instead driven by the combination of both treatments.

One strategy to minimize carryover effects, and thereby enhance the likelihood that a response is due solely to the immediately preceding treatment, is to allow sufficient time to elapse between treatments. With sufficient posttreatment delay, the participant should return to a pretreatment state before the next treatment is administered. However, it is often difficult to know how long to wait for the carryover effects to subside. Furthermore, if the delay requires the participant to return on subsequent days, attrition may become a problem.

Another essential strategy, which might be used in combination with some delay between treatments, is to counterbalance the order in which participants are exposed to the treatments. COUNTERBALANCING ensures that any carryover effects, as well as any ORDER EFFECTS (e.g., practice or fatigue effects), will fall equally on all conditions, thereby eliminating the confound between treatment and position.

—Hugh J. Foley

REFERENCES

- Girden, E. R. (1992). *ANOVA: Repeated measures* (Sage University Paper Series on Quantitative Applications in the Social Science, 07-084). Newbury Park, CA: Sage.
- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.
- Maxwell, S. E., & Delaney, H. D. (2000). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, NJ: Lawrence Erlbaum.

CART (CLASSIFICATION AND REGRESSION TREES)

The acronym CART is used in the data-mining industry both as (a) a generic for all methods of growing classification (unsupervised learning) and regression (supervised learning) trees and (b) a specific set of algorithms for growing binary trees introduced in the book by Breiman, Friedman, Olshen, and Stone (1984). The latter rely on sophisticated methods of cross-validation to adjust for problems introduced by the limitation of binary splits, which contrast with multisplit tree-growing algorithms such as CHAID (see description of CHAID).

—Jay Magidson

REFERENCE

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth.

CARTESIAN COORDINATES

The Cartesian coordinates are composed of either two or three axes (number lines) formed at 90 degrees to one another. In the two-dimensional case, x and y are often used to refer to the two axes, whereas in the three-dimensional case, a third axis, z , is added. CASES representing DATA points are located in the coordinates so that we can know the patterns of distribution of the data and the relationships between and among the variables represented by the axes.

—Tim Futing Liao

CASE

A case is the unit of study. In an aggregate analysis—say, of welfare expenditures in the American states—the case is the state, and there are 50 cases. In a SURVEY, the respondent is the case. For an EXPERIMENT, the subjects are the cases. In general, the SAMPLE size (N) indicates the number of cases, and each unit of that N represents one case.

—Michael S. Lewis-Beck

See also CASE STUDY

CASE CONTROL STUDY

The case control study is a method for estimating how various factors affect the risks of specified outcomes for large and heterogeneous populations. The method was first developed in epidemiology to identify the factors associated with heightened risks of contracting various diseases (Rothman, 1986). However, the method appears well suited for problems in the social sciences, in which researchers often aim to establish how various factors influence the risks of certain events, such as war, for a large number of countries.

Case control methods have not been exploited much in the social sciences, perhaps because their

origins in epidemiology have placed them outside most social scientists' training. This is unfortunate because case control studies have several advantages that make them particularly well suited for studying large cross-national data sets.

The method itself is quite simple. When studying rare events such as war or revolution, if one is interested in the *onset* of such conditions from prior periods of peace, then the number of such onsets is relatively few. However, years of peace will be extremely numerous. Typically, a data set of country-years (consisting of X countries observed for Y years) that includes both years of peace, as well as years in which wars or revolutions began, will have hundreds or even thousands of country-years of peace and only a few dozen country-years of event onsets. The case control method seeks to economize on data analysis by treating the thousands of peace country-years as having relatively little new information in each year. Thus, instead of pooling all country-years into the analysis—as is routinely done with most commonly used statistical methods—the case control method selects a RANDOM SAMPLE of the peace years and pools that with the event onset years. The event onset country-years are the “cases,” and the peace or nonevent years are the “controls.” A profile of independent variables is developed for each of the cases and controls, generally from the year just prior to event onset for cases and for the year selected for the controls. Conventional statistical methods such as LOGISTIC or ORDINARY LEAST SQUARES (OLS) REGRESSION can then be used to seek which variables are more strongly associated with the “cases” relative to the “controls.”

The case control method generally reduces a data set of country-year data from thousands of observations to a few hundred. However, if the controls are indeed selected randomly, the efficiency of estimating independent variable effects is virtually the same as from analysis of the entire data set (King & Zeng, 2001).

Researchers working with large cross-national data sets to study wars, revolutions, or other rare events face several major obstacles to valid inferences regarding the effects of specific variables. Two of these are practical difficulties; two others involve more substantive problems. Case control methods help with all of them.

The first practical difficulty is that large cross-national data sets are often only available for certain kinds of data, often demographic and economic factors. Although some data (e.g., on political regime type) are available for many countries and many years, other types of data (e.g., characteristics of the regime

leadership or of particular leaders) would have to be laboriously coded by hand to obtain full coverage. For other kinds of political units—such as cities, regions, or provinces—relevant data may be even more limited. Collecting or coding data for thousands of observations might be prohibitive, but doing so for a few hundred cases is often feasible. Thus, use of the case control method allows analysis of more varied and interesting variables than merely those already available with complete coverage of all countries in all years.

The second difficulty also concerns data availability. Even when data sets exist that claim to cover all countries and all years, these in fact often have extensive gaps of missing data or provide data only at 5- or 10-year intervals. Conducting conventional analyses requires filling in values for all countries and all years, thus leading to reliance on varied imputational and interpolation techniques. However, the years of interest to investigators are precisely those just prior to the onset of major disruptive events; it is in such country-years that data are most likely to be incomplete, yet routine interpolation or imputation is most likely to fail. Thus, the population of Nigeria is probably known only to within 20% and that of Canada to less than 1% accuracy. Yet to test the effects of various demographic variables on ethnic conflict, the former case is especially important. By using case control methods to reduce the number of country-years, one does not have to fill in every gap in missing data or to rely on interpolation or imputation for long stretches of data. Researchers can focus on the specific country-years drawn in the case control sample and seek to refine their accuracy on those specific observations.

There are two more substantive issues for which case control methods offer significant advantages. First, when using all country-years in conventional regression analyses, most countries will have long series of consecutive peace years. These create problems of SERIAL CORRELATION. Although econometric methods routinely counter these problems to bring out the impact of time-varying factors for individual countries, there is no good method of doing so for many different countries at once. Because case control methods use only a sample of the peace country-years, they have no such long strings of consecutive peace years in the analysis. Thus, the serial correlation problem does not arise.

Second, when using country-year data sets, using all country-years often involves comparing “apples” to “oranges.” That is, most country-years of peace

will come from rich, democratic countries, whereas most country-years of war or violence onset will come from poorer, nondemocratic ones. This might simply be because wealth and democracy lead to peace. But it might also be because peaceful but poor and non-democratic countries happen to be underrepresented in the world as we find it or because rich, democratic countries have some other characteristic (such as being clustered in Europe and North America, far from poorer areas) that conduces to peace. In any event, in using case control methods, one can help resolve such issues by matching controls with individual cases.

For example, one can stratify the random draw of controls to match the controls with the cases by year, region, colonial history, or other factors. This would increase the frequency of underrepresented groups in the data being analyzed and prevent the results from being driven by the chance distribution of countries on certain variables. Matching is particularly appropriate for variables that cannot be changed by policy but that may shift the baseline risks for groups of countries, such as region, colonial history, or time period.

Whether using simple random selection of controls or matching on certain criteria, the number of controls can be the same as the number of cases, or multiple controls can be selected for each case. What matters is having sufficient observations for significant statistical analysis, not a certain ratio of cases to controls. Thus, if one has 500 cases, a 1:1 ratio of cases to controls will provide a large enough N for most purposes; however, if one has only 100 cases, one may wish to select three or four controls for each case to achieve a larger N . Matched case control studies may require even larger numbers of controls to gain sufficient numbers of cases and controls for each stratum on the matched variable.

Epidemiology texts (e.g., Rothman 1986) provide easy guides to the estimation techniques appropriate for use with case control samples. Social scientists should make greater use of these techniques. If they can help medical analysts determine which factors cause cancer or other illnesses in a large population of individuals, they may help social scientists determine which factors cause war or other forms of violent events in a large population of varied societies.

—Jack A. Goldstone

REFERENCES

- King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9, 137–163.

Rothman, K. J. (1986). *Modern epidemiology*. Boston: Little, Brown.

CASE STUDY

The term *case study* is not used in a standard way; at face value, its usage can be misleading because there is a sense in which all research investigates cases. Nevertheless, we can identify a core meaning of the term, as referring to research that studies a small number of cases, possibly even just one, in considerable depth; although various other features are also often implied.

Case study is usually contrasted with two other influential kinds of research design: the social SURVEY and the EXPERIMENT. The contrast with the survey relates to dimensions already mentioned: the number of cases investigated and the amount of detailed information the researcher collects about each case. Other things being equal, the less cases studied, the more information can be collected about each of them. Social surveys study a large number of cases but usually gather only a relatively small amount of DATA about each one, focusing on specific features of it (cases here are usually, though not always, individual respondents). By contrast, in case study, large amounts of information are collected about one or a few cases, across a wide range of features. Here the case may be an individual (as in LIFE HISTORY work), an event, an institution, or even a whole national society or geographical region. A complication here is that when each case studied is large, social survey techniques may be used to collect data within it.

Case study can also be contrasted with experimental research. Although the latter also usually involves investigation of a small number of cases compared to survey work, what distinguishes it from case study is the fact that it involves direct control of variables. In experiments, the researcher creates the case(s) studied, whereas case study researchers identify cases out of naturally occurring social phenomena. This, too, is a dimension, not a dichotomy: QUASI-EXPERIMENTS and FIELD EXPERIMENTS occupy mid-positions.

The term *case study* is also often taken to carry implications for the *kind* of data that are collected, and perhaps also for *how these are analyzed*. Frequently, but not always, it implies the collection of unstructured data, as well as QUALITATIVE analysis of those

data. Moreover, this relates to a more fundamental issue about the purpose of the research. It is sometimes argued that the aim of case study research should be to capture cases in their uniqueness, rather than to use them as a basis for wider empirical or theoretical conclusions. This, again, is a matter of emphasis: It does not necessarily rule out an interest in coming to general conclusions, but it does imply that these are to be reached by means of inferences from what is found in particular cases, rather than through the cases being selected to test a hypothesis. In line with this, it is frequently argued that case studies should adopt an INDUCTIVE orientation, but not all case study researchers accept this position.

Another question that arises in relation to case study concerns OBJECTIVITY, in at least one sense of that term. Is the aim to produce an account of each case from an external or research point of view, one that may contradict the views of the people involved? Or is it solely to portray the character of each case “in its own terms”? This contrast is most obvious when the cases are people, so that the aim may be to “give voice” to them rather than to use them as RESPONDENTS or even as informants. However, although this distinction may seem clear-cut, in practice it is more complicated. Multiple participants involved in a case may have diverse perspectives, and even the same person may present different views on different occasions. Furthermore, there are complexities involved in determining whether what is presented can ever “capture” participant views rather than presenting an “external” gloss on them.

Some commentators point out that a case is always a “case of” something, so that an interest in some general category is built in from the beginning, even though definition of that category may change over the course of the research. In line with this, there are approaches to case study inquiry that draw on the COMPARATIVE METHOD to develop and refine theoretical categories. Examples include GROUNDED THEORIZING, historical approaches employing John Stuart Mill’s methods of agreement and difference (see Lieberman in Gomm, Hammersley, & Foster, 2000), and ANALYTIC INDUCTION.

Another area of disagreement concerns whether case study is a method—with advantages and disadvantages, which is to be used as and when appropriate, depending on the problem under investigation—or a paradigmatic approach that one simply chooses or rejects on philosophical or political grounds. Even

when viewed simply as a method, there can be variation in the specific form that case studies take:

- in the number of cases studied;
- in whether there is comparison and, if there is, in the role it plays;
- in how detailed the case studies are;
- in the size of the case(s) dealt with;
- in what researchers treat as the context of the case, how they identify it, and how much they seek to document it;
- in the extent to which case study researchers restrict themselves to description, explanation, and/or theory or engage in evaluation and/or prescription.

Variation in these respects depends to some extent on the purpose that the case study is intended to serve. When it is designed to test or illustrate a theoretical point, it will deal with the case as an instance of a type, describing it in terms of a particular theoretical framework (implicit or explicit). When it is exploratory or concerned with developing theoretical ideas, it is likely to be more detailed and open-ended in character. The same is true when the concern is with describing and/or explaining what is going on in a particular situation for its own sake. When the interest is in some problem in the situation investigated, the discussion will be geared to diagnosing that problem, identifying its sources, and perhaps outlining what can be done about it.

Variation in purpose may also inform the selection of cases for investigation. Sometimes the focus will be on cases that are judged to be typical of some category or representative of some population. Alternatively, there may be a search for deviant or extreme cases, on the grounds that these are most likely to stimulate theoretical development and/or provide a test for some hypothesis.

Many commentators, however, regard case study as more than just a method: It involves quite different assumptions about how the social world can and should be studied from those underlying other approaches (see, e.g., Hamilton, 1980). Sometimes, this is formulated in terms of a contrast between POSITIVISM, on one hand, and NATURALISM, INTERPRETIVISM, or CONSTRUCTIONISM, on the other. At the extreme, case study is viewed as more akin to the kind of portrayal of the social world that is characteristic of novelists, short story writers, and even poets. Those who see case study

in this way may regard any comparison of it with other methods in terms of advantages and disadvantages as fundamentally misconceived.

A series of methodological issues arises from these differences in view about the purpose and nature of case study and have been subjected to considerable debate:

1. *Generalizability.* In some case study work, the aim is to draw, or to provide a basis for drawing, conclusions about some general type of phenomenon or about members of a wider population of cases. A question arises here, though, as to how this is possible. Some argue that what is involved is a kind of inference or generalization that is quite different from statistical analysis, being “logical,” “theoretical,” or “analytical” in character (see Mitchell in Gomm et al., 2000; Yin, 1994). Others suggest that there are ways in which case studies can be used to make what are, in effect, the same kind of generalizations as those that survey researchers produce (see Schofield in Gomm et al., 2000). Still others argue that case studies need not make any claims about the generalizability of their findings; what is crucial is the use readers make of them—that they feed into processes of “naturalistic generalization” (see Stake in Gomm et al., 2000) or facilitate the “transfer” of findings from one setting to another on the basis of “fit” (see Lincoln & Guba in Gomm et al., 2000). There are questions, however, about whether this latter approach addresses the issue of generalizability effectively (Gomm et al., 2000, chap. 5).

2. *Causal or narrative analysis.* Case study researchers sometimes claim that by examining one or two cases, it is possible to identify CAUSAL processes in a way that is not feasible in survey research. This is because the case(s) are studied in depth, and over time rather than at a single point. It is also often argued that, by contrast with experiments, case study research can investigate causal processes “in the real world” rather than in artificially created settings. Other formulations of this argument emphasize that outcomes can always be reached by multiple pathways, so that narrative accounts of events in particular cases are essential if we are to understand those outcomes (see Becker in Gomm et al., 2000). Here, parallels may be drawn with the work of historians. In relation to all these arguments, however, there are questions about how to distinguish contingent from necessary relationships among events if only one or a small number of cases are

being studied, as well as about what role THEORY plays in causal/narrative analysis (see Gomm et al., 2000, chap. 12).

As already noted, some case study researchers argue that they can identify causal relations through comparative analysis, for example, by means of Mill's methods of agreement and difference or via analytic induction. Sometimes, the comparative method is seen as analogous to statistical analysis, but often a sharp distinction is drawn between the "logics" involved in "statistical" and "case study" work. Nevertheless, questions have been raised about whether there is any such difference in logic (Robinson in Gomm et al., 2000), as well as about the adequacy of Mill's canons and of analytic induction as means of producing theory via case study (Liebersson in Gomm et al., 2000).

3. *The nature of theory.* Although many case study researchers emphasize the role of theory, they differ in their views about the character of the theoretical perspective required. For some, it must be a theory that makes sense of the case as a bounded system. Here, the emphasis is on cases as unique configurations that can only be understood as wholes (see, e.g., Smith, 1978). For others, the task of theory is more to locate and explain what goes on within a case in terms of its wider societal context (see Burawoy, 1998). Without this, it is argued, intracase processes will be misunderstood. Indeed, it is often argued that analysis of a case always presumes some wider context; so the issue is not whether or not a macro theory is involved but rather how explicit this is and whether it is sound.

4. *Authenticity and authority.* Sometimes, case study research is advocated on the basis that it can capture the unique character of a person, situation, group, and so forth. Here there may be no concern with typicality in relation to a category or with generalizability to a population. The aim is to represent the case authentically "in its own terms." In some versions, this is seen as a basis for discovering symbolic truths of the kind that literature and art provide (see Simons, 1996). There are questions here, though, about what this involves. After all, different aesthetic theories point in divergent directions.

The commitment to authenticity may also be based on rejection of any claim to authority on the part of the case study researcher and/or on the idea that case study can be used to amplify the unique voices of those whose experience in, as well as perspective on, the world often go unheard. However, questions have been

raised about this position, not just by those committed to a scientific approach or by those who emphasize the role of macro theory, but also by some constructionists and POSTMODERNISTS. The latter's arguments undermine the notion of authenticity by denying the existence of any real phenomenon that is independent of investigations of it, by questioning the legitimacy of researchers speaking on behalf of (or even acting as mediators for) others, and/or by challenging the idea that people have unitary perspectives that are available for case study description.

In summary, although case study has grown in popularity in recent years and is of considerable value, the term is used to refer to a variety of different approaches, and it raises some fundamental methodological issues.

—Martyn Hammersley

REFERENCES

- Burawoy, M. (1998). The extended case method. *Sociological Theory*, 16(1), 4–33.
- Gomm, R., Hammersley, M., & Foster, P. (Eds.). (2000). *Case study method*. London: Sage.
- Hamilton, D. (1980). Some contrasting assumptions about case study research and survey analysis. In H. Simons (Ed.), *Towards a science of the singular: Essays about case study in educational research and evaluation* (pp. 78–92). Norwich, UK: Centre for Applied Research in Education University of East Anglia.
- Simons, H. (1996). The paradox of case study. *Cambridge Journal of Education*, 26(2), 225–240.
- Smith, L. M. (1978). An evolving logic of participant observation, educational ethnography and other case studies. *Review of Research in Education*, 6, 316–377.
- Yin, R. (1994). *Case study research* (2nd ed.). Thousand Oaks, CA: Sage.

CASI. See COMPUTER-ASSISTED DATA COLLECTION

CATASTROPHE THEORY

Catastrophe theory refers to a type of behavior among some nonlinear dynamic mathematical models that experience NONLINEAR DYNAMICS such that sudden or rapid large-magnitude changes in the value of one

variable are a consequence of a small change that occurs in the value of a parameter (called a *control parameter*). In this sense, catastrophe theory can model phenomena that loosely follow a “straw that broke the camel’s back” scenario, although catastrophe theory can be very general in its application. The modern understanding of catastrophe theory has its genesis in work by Thom (1975).

Nearly all early work with catastrophe theory employed polynomial functions in the specification of differential equation mathematical models. In part, this was an important consequence of the generality of Thom’s (1975) findings. Because all sufficiently smooth functions can be expanded using a Taylor series approximation (which leads us to a polynomial representation of the original model), it is possible to analyze the polynomial representation directly (see Saunders, 1980, p. 20). However, scientists can avoid using one of Thom’s canonical polynomial forms by working with their own original theory-rich specifications as long as it is clear that the original specification has catastrophe potential (Brown, 1995a).

Fundamental to catastrophe theory is the idea of a bifurcation. A bifurcation is an event that occurs in the evolution of a dynamic system in which the characteristic behavior of the system is transformed. This occurs when an attractor in the system changes in response to change in the value of a parameter (called a *control parameter* because its value controls the manifestation of the catastrophe). A catastrophe is one type of bifurcation (as compared with, say, subtle bifurcations or explosive bifurcations). The characteristic behavior of a dynamic system is determined by the behavior of trajectories, which are the values of the variables in a system as they change over time. When trajectories intersect with a bifurcation, they typically assume a radically different type of behavior as compared with what occurred prior to the impact with the bifurcation. Thus, if a trajectory is “hugging” close to an attractor or equilibrium point in a system and then intersects with a bifurcation point, the trajectory may suddenly abandon the previous attractor and “fly” into the neighborhood of a different attractor. The fundamental characteristic of a catastrophe is the sudden disappearance of one attractor and its basin, combined with the dominant emergence of another attractor. Because multidimensional surfaces can also attract (together with attracting points on these surfaces), these gravity centers within dynamical systems are referenced more generally as attracting hypersurfaces, limit sets, or simply attractors.

The following model, which is due to Zeeman (1972), illustrates a simple cusp catastrophe and is used to model the change in muscle fiber length (variable x) in a beating heart. The control parameter A (which in this instance refers to the electrochemical activity that ultimately instructs the heart when to beat) can change in its value continuously, and it is used to move trajectories across an equilibrium hypersurface that has catastrophe potential. The parameter q identifies the overall tension in the system, and f is a scaling parameter. The two differential equations in this system are $dx/dt = -f(x^3 - qx + A)$ and $dA/dt = x - x_1$. Here, x_1 represents the muscle fiber length at systole (the contracted heart equilibrium). Setting the derivative $dx/dt = 0$, we will find between one and three values for x , depending on the other values of the system. When there are three equilibria for x for a given value of the control parameter A , one of the equilibria is unstable and does not attract any trajectory. The other two equilibria compete for the attention of the surrounding trajectories, and when a trajectory passes a bifurcation point in the system, the trajectory abandons one of these equilibria and quickly repositions itself into the neighborhood (i.e., the basin) of the other equilibrium. This rapid repositioning of the trajectory is the catastrophe.

Two social scientific examples of nonlinear differential equation dynamic catastrophe specifications have been developed and explored by Brown (1995b, chaps. 3 and 5). One example involves the interaction between candidate preferences, feelings for a political party, and the quality of an individual’s political context or milieu during the 1980 presidential election in the United States. The other example addresses the interaction between the partisan fragmentation of the Weimar Republic’s electorate, electoral de-institutionalization, and support for the Nazi party. Both examples are fully estimated; the first uses both individual and aggregate data, whereas the second employs aggregate data only.

—Courtney Brown

REFERENCES

- Brown, C. (1995a). *Chaos and catastrophe theories*. Thousand Oaks, CA: Sage.
- Brown, C. (1995b). *Serpents in the sand: Essays on the nonlinear nature of politics and human destiny*. Ann Arbor: University of Michigan Press.
- Saunders, P. T. (1980). *An introduction to catastrophe theory*. New York: Cambridge University Press.
- Thom, R. (1975). *Structural stability and morphogenesis*. Reading, MA: Benjamin.

Zeeman, E. C. (1972). Differential equations for the heartbeat and nerve impulse. In C. H. Waddington (Ed.), *Towards a theoretical biology* (Vol. 4, pp. 8–67). Edinburgh, UK: Edinburgh University Press.

CATEGORICAL

Categorical is a term that includes such terms as NOMINAL VARIABLE and ordinal variable. It derives its name from the fact that, in essence, a categorical variable is simply a categorization of observations into discrete groupings, such as the religious affiliations of a sample.

—Alan Bryman

See also ATTRIBUTE, DISCRETE, NOMINAL VARIABLE

CATEGORICAL DATA ANALYSIS

Social science data, when quantified, can be subject to a variety of statistical analyses, the most common of which is regression analysis. However, the requirements of a continuous dependent variable and of other regression assumptions make linear regression sometimes a less desirable analytic tool because a lot of social science data are categorical.

Social science data can be categorical in two common ways. First, a variable is categorical when it records nominal or discrete groups. In political science, a political candidate in the United States can be a Democrat, Republican, or Independent. In sociology, one's occupation is studied as a discrete outcome. In demography, one's contraceptive choice such as the pill or the condom is categorical. Education researchers may study discrete higher education objectives of high school seniors: university, community college, or vocational school.

Furthermore, a variable can take on an ordinal scale such as the LIKERT SCALE or a scale that resembles it. Such a scale is widely used in psychology in particular and in the social sciences in general for measuring personality traits, attitudes, and opinions and typically has five ordered categories ranging from *most important* to *least important* or from *strongly agree* to *strongly disagree* with a neutral middle category. An ordinal scale has two major features: There exists a natural order among the categories, and the

distance between a lower positioned category and the next category in the scale is not necessarily evenly distributed throughout the scale. Variables not measuring personality or attitudes can also be represented with an ordinal scale. For example, one's educational attainment can be simply measured with the three categories of "primary," "secondary," and "higher education." These categories follow a natural ordering, but the distance between "primary" and "secondary" and that between "secondary" and "higher education" is not necessarily equal. It is apparent that an ordinal variable has at least three ordered categories. Although there is no upper limit for the total number of categories, researchers seldom use a scale of more than seven ordered categories.

The examples of categorical data above illustrate their prevalence in the social sciences. Categorical data analysis, in practice, is the analysis of categorical *response* variables.

HISTORICAL DEVELOPMENT

The work by Karl Pearson and G. Udny Yule on the association between categorical variables at the turn of the 20th century paved the way for later development in models for discrete responses. Pearson's contribution is well known through his namesake chi-square statistic, whereas Yule was a strong proponent of the ODDS RATIO in analyzing association. However, despite important contributions by noted statisticians such as R. A. Fisher and William Cochran, categorical data analysis as we know it today did not develop until the 1960s.

The postwar decades saw a rising interest in explaining social issues and a burgeoning need for skilled social science researchers. As North American universities increased in size to accommodate postwar baby boomers, so did their faculties and the bodies of university-based social science researchers. Categorical scales come naturally for measuring attitudes, social class, and many other attributes and concepts in the social sciences. Increasingly in the 1960s, social surveys were conducted, and otherwise quantifiable data were obtained. The increasing methodological sophistication in the social sciences satisfied the increasing need for analytic methods for handling the increasingly available categorical data. It is not surprising that many major statisticians who developed regression-type models for discrete responses were all academicians with social sciences affiliations

or ties, such as Leo Goodman, Shelby Haberman, Frederick Mosteller, Stephen Fienberg, and Clifford Clogg. These methodologists focused on log-linear models, whereas their counterparts in the biomedical sciences concentrated their research on regression-type models for categorical data. Together, they and their biomedical colleagues have taken categorical data analysis to a new level.

The two decades before the 21st century witnessed a greater concern with models studying the association between ordinal variables, especially when their values are assumed to be latent or not directly measurable. The same period also saw a resurgence of interest in latent variable models studied by Lazarsfeld in the 1950s (e.g., latent class analysis) and their extension to regression-type settings. At the forefront of these two areas were Leo Goodman, Clifford Clogg, and Jacque Hagenaars, who represented the best of both North American and European social science methodologists.

Another tradition in categorical data analysis stems from the concerns in econometrics with a regression-type model with a dependent variable that is limited in one way or another (in observed values or in distribution), a notion popularized by G. S. Maddala. When only two values (binary) or a limited number of ordered (ordinal) or unordered (nominal) categories can be observed for the dependent variable, LOGIT and PROBIT models are in order. During the past quarter century, computer software has mushroomed to implement these binary, ordinal, and multinomial logit or probit models. In statistics, many models for categorical data such as the logit, the probit, and Poisson regression as well as linear regression can be more conveniently represented and studied as the family of GENERALIZED LINEAR MODELS.

TYPES OF CATEGORICAL DATA ANALYSIS

There are several ways to classify categorical data analysis. Depending on the parametric nature of a method, two types of categorical data analysis arise.

1. *Nonparametric methods.* These methods make minimal assumptions and are useful for hypothesis testing. Examples include the Pearson CHI-SQUARE TEST, Fisher's exact test (for small expected frequencies), and the Mantel-Haenszel test (for ordered categories in linear association).

2. *Parametric methods.* This model-based approach assumes random (possibly stratified) sampling and

is useful for estimation purposes and flexible for fitting many specialized models (e.g., symmetry, quasi-symmetry). It also allows the estimation of the STANDARD ERROR, COVARIANCE, and CONFIDENCE INTERVAL of model parameters as well as predicted response probability.

The second type is the most popular in the social sciences, evidenced by the wide application of parametric models for categorical data. Variables in social science research can occupy dependent or independent positions in a statistical model. They are also known as response and explanatory variables, respectively. The type (categorical or not) and the positioning of a variable (DEPENDENT or INDEPENDENT) give rise to the statistical model(s) for (categorical) data, providing another way to classify categorical data analysis. For simplicity, we view all variables as either categorical or continuous (i.e., with interval or ratio scales). In some models, all variables can be considered dependent because only the association among them is of interest to the researcher. For the models that distinguish between dependent and independent variables, we limit our attention to those with single dependent variables. That is, in such models, there is only one dependent variable per model regardless of the number of independent variables, which can be all continuous, all categorical, or a mixture of both.

Models of association do not make a distinction between response and explanatory variables in the sense that the variables under investigation depend on each other. When both variables in a pair are continuous, we use correlation analysis; when they are categorical, we use contingency table and log-linear analysis. The situation can be generalized to multiple variables as in the case of partial correlation analysis and multiway contingency tables.

When the two variables in a pair are of different types (i.e., one categorical and the other continuous), a causal order is assumed with the separation of the independent and the dependent variables. For analyzing the distribution of a continuous-response variable in explanatory categories, we apply ANOVA or linear regression, which can be extended into including additional explanatory variables that are continuous. In that case, we use ANCOVA and linear regression. The flexibility of logit and probit models is evidenced by their applicability to all three cells of Table 1 for a categorical response variable, regardless of the type of explanatory variables—continuous, categorical, or mixed. Often

Table 1 Correspondence Between Variable Types and Statistical Models

		<i>Dependent</i>	
		<i>Continuous</i>	<i>Categorical</i>
Dependent	Continuous	Correlation	
	Categorical		Contingency table analysis, log-linear models
Independent	Continuous	Linear regression	Logit and probit models
	Categorical	Continuous	Logit and probit models
	Mixed	ANCOVA, linear regression	Logit and probit models

Table 2 Types of Latent Variable Models According to the Type of Variables

	<i>Latent Continuous</i>	<i>Latent Categorical</i>
Observed continuous	Factor-analytic model; SEM	Latent profile model
Observed categorical	Latent trait model; SEM	Latent class model

the same data may be analyzed using different methods, depending on the assumptions the researcher is willing to make and the purposes of the research.

Thus, in Table 1, the cells in the column under the heading of categorical dependent variables comprise the methods of categorical data analysis. A further consideration of these methods is whether the response variable has ordered categories as opposed to pure nominal ones. Excluded from the table, for instance, is Poisson regression that models an integer-dependent variable following a Poisson distribution, which is also a limited-dependent variable model.

Depending on whether a variable is discrete or ordinal, a model for categorical data can be further classified. For example, a logit (or probit) model is binary when the dependent variable is dichotomous, is ordinal (or ordered) when the dependent variable has ordered categories, and is multinomial when it has more than two discrete categories. Log-linear models for the association between ordinal variables can be extended into linear-by-linear association and log-multiplicative models, to name just two examples (see LOG-LINEAR MODEL and ASSOCIATION MODEL for further details).

We have so far examined categorical data analysis with observed variables only. Social scientists also employ in categorical data analysis the LATENT VARIABLE to represent concepts that cannot be measured directly. Here we have yet another way of

classifying categorical data analysis, depending on whether the latent variables, the observed variables, or both in a model are categorical.

Specifically, categorical data analysis is concerned with the second row in Table 2, in which the observed variables are categorical. LATENT CLASS ANALYSIS, which estimates categorical latent classes using categorical observed variables, has received most attention in the social sciences.

Several major computer programs have made the latent class model much more accessible by many social scientists. These include C. C. Clogg's mainframe program MLLSA, which is currently available in S. Eliason's Categorical Data Analysis System at the University of Minnesota; S. Haberman's NEWTON and DNEWTON; J. Hagenaars's LCAG; and J. Vermunt's LEM, among others.

MODEL FITTING AND TEST STATISTICS

Yet one more way to view and classify categorical data analysis looks at how model fitting is defined. Much of current social science research using categorical data derives from three general yet related traditions: one that fits EXPECTED FREQUENCIES to observed frequencies, another that fits expected response values to observed response values, and a last that fits expected measures of association through correlation to the observed counterparts. The third tradition describes how categorical data are handled in factor-analytic and structural equation modeling approaches, and it receives separate attention elsewhere in the encyclopedia. We focus here on the first two traditions.

The first tradition fits a particular statistical model to a table of frequencies. Let f_i indicate observed frequency in cell i and F_i indicate expected frequency in cell i in the table. Then our purpose in model fitting is to reduce some form of the difference between the

observed and the expected frequencies in the table. Thus, we obtain the Pearson chi-square statistic and the LIKELIHOOD RATIO STATISTIC L^2 (or G^2):

$$\chi^2 = \sum_i \frac{(f_i - F_i)^2}{F_i} \quad \text{and} \quad L^2 = 2 \sum_i f_i \log\left(\frac{f_i}{F_i}\right).$$

The summation is over all cell i except the ones with structural zeros. These statistics can be used alone in a nonparametric method for a table of frequencies or in a parametric method such as a log-linear model, which typically expresses $\log(F_i)$ as a linear combination of the effects of the rows, columns, layers, and so on and their interactions in a table. The two test statistics given above are actually special cases of a family of test statistics known as power divergence statistics, which includes a third member, the CRESSIE-READ STATISTIC, that represents a compromise between χ^2 and L^2 .

The second tradition is that of the generalized linear model, which attempts to model the expected value μ of the response variable y as a function of a linear combination of the explanatory variables and their parameters. The data are expected to follow a form of exponential distribution, and the function linking $g(\mu)$ to the linear combination of the independent variables and parameters may take on various forms, such as the widely used function of the logit.

The two traditions are closely related. For example, when all variables—response and explanatory—are categorical, $F_i = n_i \mu_i$, where n_i is the number of observations in cell i . The test statistics described above also apply to the assessment of model fitting of generalized linear models.

MODEL COMPARISON AND SELECTION

This is a prominent issue in categorical data analysis. Researchers often feel the need to compare statistical models and to choose a better fitting model from a set of models. The latter situation describes hierarchical log-linear model fitting when a number of log-linear models embedded in one another are considered. These models have different degrees of freedom because they include a different number of parameters. Differences in L^2 give comparative likelihood ratio statistics for drawing conclusions.

Sometimes researchers need to compare models that are nonhierarchical. This can be a comparison of the model fitting of two (or more) subgroups of observations or two (or more) models that include different

sets of variables. Although sometimes treating these subgroups or submodels as members of a single combined supermodel is a useful exercise, this approach is not always realistic because the submodels under comparison can be just incompatible. Therefore, a statistical criterion of a model's goodness of fit that applies across different models with different degrees of freedom is necessary. Two widely applied criteria belonging to this category are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

Relying on information-theoretic considerations, the AIC adjusts the likelihood ratio statistic with twice the number of degrees of freedom in a given model:

$$\text{AIC} = L^2 - 2(df),$$

where df represents the number of degrees of freedom in the model. Between two models, hierarchical or not, the model with the smaller AIC value is the better one.

Based on Bayesian decision theory, the BIC further adjusts the likelihood ratio statistic with the total sample size (N):

$$\text{BIC} = L^2 - \log(N)(df).$$

The BIC gives an approximation of the logarithm of the Bayes factor for model comparison. It has been shown that the BIC penalizes the model with a larger degree of freedom more so than the AIC if sample sizes are not very small. When comparing hierarchical models in which L^2 differences are also applicable, the BIC usually favors a parsimonious model more so than the other goodness-of-fit statistics.

MISSING DATA IN CATEGORICAL DATA ANALYSIS

MISSING DATA can be a serious concern in any multivariate analysis and deserve particular attention in models of frequency tables that may have zero cells. Of special interest are the so-called partially observed data. If an observation has missing information on all variables, the observation cannot be used at all. But if an observation has missing information on some of the variables in the model, the conventional approach of listwise deletion would waste useful information. Added to this problem is the mechanism with which missing information occurs. Of the three mechanisms of missing data—missing completely at random, missing at random, and missing

not at random—the last may create bias in drawing inference about relationships between variables in the model. Developments in recent years have sought to deal with estimating statistical models of categorical data under various missing data mechanisms. The LEM software by J. Vermunt is the most flexible for such purposes.

—Tim Futing Liao

REFERENCES

- Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley.
- Clogg, C. C., & Shihadeh, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.
- Everitt, B. S. (1992). *The analysis of contingency tables*. London: Chapman & Hall.
- Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge: MIT Press.
- Liao, T. F. (1994). *Interpreting probability models: Logit, probit, and other generalized linear models*. Thousand Oaks, CA: Sage.
- Long, S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Powers, D. A., & Xie, Y. (2000). *Statistical models for categorical data analysis*. New York: Academic Press.

CATI. See COMPUTER-ASSISTED DATA COLLECTION

CAUSAL MECHANISMS

Causal mechanisms involve the processes or pathways through which an outcome is brought into being. We explain an outcome by offering a HYPOTHESIS about the cause(s) that typically bring it about. So a central ambition of virtually all social research is to discover causes. Consider an example: A rise in prices causes a reduction in consumption. The causal mechanism linking cause to effect involves the choices of the rational consumers who observe the price rise, adjust their consumption to maximize overall utility, and reduce their individual consumption of this good. In the aggregate, this rational behavior at the individual level produces the effect of lower aggregate consumption.

There are two broad types of theories of causation: the Humean theory (“causation as regularities”)

and the causal realist theory (“causation as causal mechanism”). The Humean theory holds that causation is entirely constituted by facts about empirical regularities among observable variables; there is no underlying causal nature, causal power, or causal necessity. The causal realist takes notions of causal mechanisms and causal powers as fundamental and holds that the task of scientific research is to arrive at empirically justified THEORIES and hypotheses about those causal mechanisms. Consider these various assertions about the statement, “X caused Y”:

- X is a necessary and/or sufficient condition of Y.
- If X had not occurred, Y would not have occurred.
- The conditional probability of Y given X is different from the absolute probability of Y [$P(Y|X) \neq P(Y)$].
- X appears with a nonzero coefficient in a REGRESSION equation PREDICTING the value of Y.
- There is a causal mechanism leading from the occurrence of X to the occurrence of Y.

The central insight of causal realism is that the final criterion is in fact the most fundamental. According to causal realism, the fact of the existence of underlying causal mechanisms linking X to Y accounts for each of the other criteria; the other criteria are symptoms of the fact that there is a causal pathway linking X to Y.

Causal reasoning thus presupposes the presence of a causal mechanism; the researcher ought to attempt to identify the unseen causal mechanism joining the variables of interest. The following list of causal criteria suggests a variety of ways of using available evidence to test or confirm a causal hypothesis: apply Mill’s methods of similarity and difference as a test for necessary and sufficient conditions, examine conditional probabilities, examine CORRELATIONS and regressions among the variables of interest, and use appropriate parts of social theory to hypothesize about underlying causal mechanisms. Causal realism insists, finally, that EMPIRICAL evidence must be advanced to assess the credibility of the causal mechanism that is postulated between cause and effect.

What is a causal mechanism? A causal mechanism is a sequence of events or conditions, governed by lawlike regularities, leading from the explanans to the explanandum. Wesley Salmon (1984) puts the point

this way: “Causal processes, causal interactions, and causal laws provide the mechanisms by which the world works; to understand *why* certain things happen, we need to see *how* they are produced by these mechanisms” (p. 132). Nancy Cartwright likewise places real causal mechanisms at the center of her account of scientific knowledge. As she and John Dupré put the point, “Things and events have causal capacities: in virtue of the properties they possess, they have the power to bring about other events or states” (Dupré & Cartwright, 1988). Most fundamentally, Cartwright argues that identifying causal relations requires substantive theories of the causal powers or capacities that govern the entities in question. Causal relations cannot be directly inferred from facts about association among variables.

The general nature of the mechanisms that underlie social causation has been the subject of debate. Several broad approaches may be identified: agent-based models, structural models, and social influence models. Agent-based models follow the strategy of aggregating the results of individual-level choices into macro-level outcomes, structural models attempt to demonstrate the causal effects of given social structures or institutions (e.g., the tax collection system) on social outcomes (levels of compliance), and social influence models attempt to identify the factors that work behind the backs of agents to influence their choices. Thomas Schelling’s (1978) apt title, *Micromotives and Macrobehavior*, captures the logic of the former approach, and his work profoundly illustrates the sometimes highly unpredictable results of the interactions of locally rational-intentional behavior. Jon Elster has also shed light on the ways in which the tools of rational choice theory support the construction of large-scale sociological explanations (Elster, 1989). Emirbayer and Mische (1998) provide an extensive review of the current state of debate on the concept of agency. Structuralist and social influence approaches attempt to identify socially salient influences such as institution, state, race, gender, and educational status and provide detailed accounts of how these factors influence or constrain individual trajectories, thereby affecting social outcomes.

—Daniel Little

REFERENCES

- Cartwright, N. (1989). *Nature’s capacities and their measurement*. Oxford, UK: Oxford University Press.
- Dupré, J., & Cartwright, N. (1988). Probability and causality: Why Hume and indeterminism don’t mix. *Nous*, 22, 521–536.
- Elster, J. (1989). *The cement of society: A study of social order*. Cambridge, UK: Cambridge University Press.
- Emirbayer, M., & Mische, A. (1998). What is agency? *American Journal of Sociology*, 103(4), 962–1023.
- Little, D. (1998). *Microfoundations, method and causation: On the philosophy of the social sciences*. New Brunswick, NJ: Transaction Publishers.
- Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford, UK: Clarendon.
- Miller, R. W. (1987). *Fact and method: Explanation, confirmation and reality in the natural and the social sciences*. Princeton, NJ: Princeton University Press.
- Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.
- Schelling, T. C. (1978). *Micromotives and macrobehavior*. New York: Norton.

CAUSAL MODELING

In the analysis of social science data, researchers to a great extent rely on methods to infer causal relationships between CONCEPTS represented by measured VARIABLES—notably, those between DEPENDENT VARIABLES and INDEPENDENT VARIABLES. Although the issue of CAUSALITY can be rather philosophical, a number of analytical approaches exist to assist the researcher in understanding and estimating causal relationships. Causal thinking underlies the basic design of a range of methods as diverse as EXPERIMENTS, and models based on the MARKOV CHAIN. One thing that distinguishes experimental from nonexperimental design is the fact that the independent variables are actually set at fixed values, or manipulated, rather than merely observed. Thus, with nonexperimental or observational design, causal inference is much weaker. Given an observational design, certain conditions can strengthen the argument that the finding of a relationship between, say, *X* and *Y* is causal rather than SPURIOUS. Of particular importance is temporality (i.e., a demonstration that *X* occurred before *Y* in time). Also important is meeting the ASSUMPTIONS necessary for causal inference. In classical REGRESSION analysis, these are sometimes referred to as the GAUSS-MARKOV assumptions.

Methods for analyzing observational data also include all kinds of regression-type models subsumed

under GENERAL LINEAR MODELS or GENERALIZED LINEAR MODELS, especially GRAPHIC MODELING, PATH ANALYSIS, or STRUCTURAL EQUATION MODELING (SEM), which allow for multiequation or SIMULTANEOUS EQUATION systems. SEM, which can include LATENT VARIABLES to better represent CONCEPTS, can be estimated with software such as AMOS, EQS, LISREL, or M-PLUS.

A typical simultaneous equation system may contain RECIPROCAL RELATIONSHIPS (e.g., X causes Y and Y causes X), which renders them NONRECURSIVE. Here is an example of such a system, which may be called a causal model:

$$\begin{aligned} Y_1 &= b_{10} + b_{11}X_1 + b_{12}Y_2 + e_1, \\ Y_2 &= b_{20} + b_{21}Y_1 + b_{22}X_2 + e_2. \end{aligned}$$

The phrase *causal model* is taken to be a synonym for a system of simultaneous, or structural, equations. Sometimes, too, a sketched arrow diagram, commonly called a *path diagram*, accompanies the presentation of the system of equations itself. Before estimation proceeds, the IDENTIFICATION PROBLEM must be solved through examination of the relative number of EXOGENOUS and ENDOGENOUS variables to see if the order and rank conditions are met. In this causal model, the X variables are exogenous, the Y variables endogenous. As it turns out, when each equation in this system is exactly identified, the entire system is identified, and its parameters can be estimated. By convention, exogenous variables are labeled X and endogenous variables labeled Y , but it must be emphasized that simply labeling them such does not automatically make them so. Instead, the assumptions of exogeneity and endogeneity must be met.

Once the model is identified, estimation usually goes forward with some INSTRUMENTAL VARIABLES method, such as two-stage least squares. A three-stage least squares procedure also exists, but it has not generally been shown to offer any real improvement over the two-stage technique. The method known as indirect least squares can be employed if each equation in the system is exactly identified. ORDINARY LEAST SQUARES cannot be used for it will generally produce biased and inconsistent estimates unless the system is RECURSIVE, whereby causality is one-way and the error terms are uncorrelated across equations. There are MAXIMUM LIKELIHOOD estimation analogs to two-stage and three-stage least squares: limited-information maximum likelihood and full-information maximum likelihood, respectively.

Causal modeling, as the discussion so far suggests, relies on statistical analysis that has PROBABILITY theory as its foundation. Causality describes a process imbued in uncertainty, and it is the job of the analyst to take into account the uncertainty while capturing the causal patterns.

—Michael S. Lewis-Beck and Tim Futing Liao

REFERENCES

- Berry, W. (1984). *Nonrecursive causal models*. Beverly Hills, CA: Sage.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley.
- Davis, J. (1985). *The logic of causal order*. Beverly Hills, CA: Sage.
- Kmenta, J. (1997). *Elements of econometrics* (2nd ed.). Ann Arbor: University of Michigan Press.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

CAUSALITY

Often a hesitation is expressed when causality stands central in social science discussions and debates. “We are afraid of causality,” so it sounds. But the discussants all admit that they cannot do without causality: “In the end, we need causal explanations.” This hesitation also holds for the notion of causality in the history of philosophy. The historical train of causality is like a local train; it resembles a procession of *Echternach*, going now two steps forward and then one step backward. Aristotle, the schoolmen, Leibniz, Kant, and Schopenhauer were causalists. But Hume, Mach, Pearson, Reichenbach, the Copenhagen school of quantum physicists, and Russell were a-causalists. It was Bertrand Russell (1912) who left us with the following bold expression: “The law of causality, I believe, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm” (p. 171).

ARISTOTLE: THE FOUNDING FATHER

We may consider Aristotle to be the founding father of causal thinking. He was a philosopher who distinguished many types of causes (e.g., *causa materialis*, *causa formalis*, *causa efficiens*, and *causa finalis*). In modern science of the 17th century, his *causa efficiens* or labor cause became the guiding if not the

only principle. Causation was then associated with “production.” The labor cause was the active agent, the external power was like the knock with a hammer, and the effect was that which undergoes in a passive way, like the chestnut that bursts into pieces. This notion of production is present in Newton’s external power, which brings a resting body into movement. It is also present in social policymaking. For example, policymakers in France in the 19th century started a policy of raising the birth rate and of encouraging immigration to establish an excess of births over deaths, so that the aging population structure could be reversed. It is also present in the stimulus-response scheme of the school of BEHAVIORISM. And it is even present in our language because many verbs such as *poison*, *cure*, *calm*, *humidify*, and *illuminate* contain the idea of production.

GALILEI: NECESSARY AND SUFFICIENT CONDITION

It was also in modern science that cause was clarified in terms of a necessary and sufficient condition by Galilei, where *necessary* means “*conditio sine qua non*” or “If not *X*, then not *Y*,” and *sufficient* means “If *X*, then always *Y*.” (One should notice that “*X* necessitates *Y*” means “*X* is a sufficient condition of *Y*.” Confusion is possible here!)

DAVID HUME: A TURNING POINT IN HISTORY

David Hume in the 18th century is really a turning point in the history of causality. In fact, no author today can still write about this subject without first discussing Hume. An outstanding example is Karl Popper (1972), who, in his book *Objective Knowledge*, has a first chapter titled “The Hume Problem.” Hume was not against causality (see Hume, 1786). He believed that the world was full of it. But he was skeptical about the possibility of science getting insight into the question of why a cause is followed by an effect. His starting point is purely *empirical*. He argued that knowledge of causal relations was never brought about in an *a priori* fashion by means of pure deduction but that it was totally based on experience. Adam could not deduce from seeing water that it could suffocate him. Hume also believed that causes were external to their effects. If a billiard ball collides with a second ball and the latter starts moving, then there is nothing present in the first ball that gives us the slightest idea about

what is going to happen to the second one. As for the middle term in between cause and effect (i.e., the causal arrow), Hume stated that such concepts as production, energy, power, and so forth belonged to an obscure philosophy that served as a shelter for superstition and as a cloak for covering foolishness and errors. We see the fire and feel the heat, but we cannot even guess or imagine the connection between the two. We are not even directly conscious of the energy with which our will influences the organs of our bodies, such that it will always escape our diligent investigations. The question of why the will influences the tongue and the fingers, but not the heart and the liver, will always bring us embarrassment. And the idea that the will of a Supreme Being is responsible here brings us far beyond the limits of our capabilities. Our perpendicular line is too short to plumb these yawning chasms. He concluded that when we saw both the cause and the effect, then there would be *constant conjunction*. After some time of getting used to this, *custom* arises. And then, via some mechanism of *psychological association*, we gradually start to develop a *belief*. And on the basis of this belief, we use causal terminology and make *predictions*. In short, the only thing that really exists is constant conjunction (regularity theory); the rest is psychology. Loosely speaking, correlation is the matter; the rest is chatter.

IMMANUEL KANT: CAUSALITY IS AN A PRIORI CATEGORY

With the reactions on David Hume, one can fill a whole library. Among others, Immanuel Kant, who first admitted that Hume had awakened him out of his dogmatic half-sleep, refused to accept that only experience is the basis of causality (see Kant, 1781). He believed in prior categories of understanding, which, together with experience, brought us the synthesis called objective knowledge. Causality was, in his view, one of these categories of understanding. This view of causality as an *a priori* category is not undisputed. John Mackie (1974) gave the example of a piece of wood that is cut into two parts and argued that from a long enough distance, we would first see the parts fall apart and then hear the sound of the axe due to the difference in the speed of the light and the sound, but up close, we would first hear it and then see the parts fall apart due to the inertia of the bodies. And if a little train on the table bumps into a second train, whereby the second train starts moving, then this has

nothing to do with our prior faculty of the mind because there might be a hidden magnet under the table that brings about the second movement. So, there is not much of a priori causal knowledge here but rather the act of hypothesizing knowledge in a tentative way, with experience and observation needed to come to a conclusion.

MODERN SCIENCE: HYPOTHESIS TESTING

The tradition of hypothesis testing, often used today in our scientific research, was initiated by positivist philosophers of the 19th century such as August Comte and John Stuart Mill and became standard procedure within positivist research in the 20th century. Karl Popper (1959) and Carl Gustav Hempel (1965) also contributed to this general attitude by means of their deductive-nomological approach, which is well known as the covering law theory (i.e., every concrete causal statement is covered by general laws that operate in the background and serve as general theories and hypotheses).

THE PROBABILISTIC THEORY OF CAUSALITY

In the meantime, the debate on causality goes on. Most proposals contain a policy of encirclement because causal production is not approached in a direct way but rather indirectly via some other criteria. One such criterion is “probability,” used by a school of scientists who may be called the adherents of *A Probabilistic Theory of Causality*. This is the title of a book by Patrick Suppes, who initiated this school of thought in 1970. His opinion corresponds with the opinion of the majority of scientific researchers in the world—that is, X is a cause of Y if and only if X exists (the probability of X is greater than zero), X is temporally prior to Y (the moment in time t_X comes before t_Y), there is a statistical relationship between X and Y (the probability of Y given X is greater than the probability of Y by itself), and there is no spuriousness (the statistical relationship between X and Y does not disappear when controlling other potentially CONFOUNDING factors). These criteria for causality are in use in social research, especially in statistical analysis. The method of path coefficients of Sewell Wright in genetics in the 1920s, the simultaneous equation models of Herman Wold in econometrics in the 1950s, the causal models of Simon and Blalock in sociology and other social sciences in the 1960s and 1970s, and the linear

structural relations system (LISREL) of Karl Jöreskog in the 1970s and after are but a few of the many examples. Three main problems will always haunt this school of thought. The first is the notion of probability, which explains the statistical relationship (i.e., correlation) but not causation. The second is the dependence on theory, which is expressed by the relations between the variables in the causal model, especially by the factors to control for spuriousness. The third problem is that temporal priority is used instead of causal priority, which means that research practice is building on the sophism “Post hoc, ergo propter hoc” (i.e., Thereafter, so therefore).

MARIO BUNGE: PARADISE OF SIMPLISSIMUS

Another policy of encirclement is the use of Galilei’s criterion of a “necessary” and “sufficient” condition, which is done by many modern authors but in the most intelligent way by John Mackie (1974). He reacts against the extreme standpoint of Mario Bunge (1959), who tells us that we commit the sin of meaning-inflation if we liberalize the concept of causality so that it covers nearly everything. In his view, there are many forms of determination in reality—probabilistic, dialectical, functional, structural, mechanical, and other determinations—of which the causal determination is only one. The core of this causal determination is the criterion given by Aristotle’s *causa efficiens* (i.e., necessary production). Next to productive, the causal relation is also conditional (if X , then Y), unique (one cause, one effect), asymmetrical (when X causes Y , then Y does not cause X), and invariable (no exceptions, no probabilistic causality). Bunge also gives other strict criteria, such as linearity, additivity, continuity, and the like. His notion of causality is very rigid. He reacts against functionalists, interactionists, and dialecticians, who see everything in terms of interdependence and are, therefore, romanticists. In their view, causality is a key to every door, a panacea. In Bunge’s view, causal relations, if strictly defined, are only a small part of reality, but they exist. They are neither myth nor panacea.

JOHN MACKIE: INUS CONDITION IN A CAUSAL FIELD

It goes without saying that Bunge’s (1959) rigid definition of causality is a paradise of simplicity. John Mackie (1974) reacts against this rigidity in

developing a very liberal notion of causality that is close to everyday language and accepts probabilism, multicausality, teleology, functionalism, and many other forms of determination. His approach is in terms of necessary and sufficient condition. A causal factor is, in his view, a necessary condition “in the circumstances,” which means that silent premises and particulars should, as much as possible, be made explicit. For example, a fire breaks out in a house. Experts claim that this is due to a short circuit. They do not give a necessary condition because other things than a short circuit, such as the falling of an oil heater, could have caused the fire. They do not give a sufficient condition because even if there was a short circuit, other conditions such as inflammable materials or the absence of an efficient fire extinguisher are necessary for a fire to start. Therefore, there is a complex set of conditions, positive and negative, that together with a short circuit are sufficient but not necessary for the outbreak of a fire because other factors could have caused the fire. A short circuit is a necessary part of the set of conditions, for without it, the inflammable materials and absence of an efficient fire extinguisher could not cause a fire. A short circuit is then an insufficient but necessary part of a complex set of conditions, of which the total is unnecessary but sufficient for the result. In short, a short circuit is an INUS condition for fire—that is, an Insufficient but Necessary part of a set, which is Unnecessary but Sufficient for the result. Mackie adds to this that there will always be factors that cannot vary but are fixed in the causal field. For example, being born is, in the strict logical sense, an INUS condition of dying, but it cannot be a candidate cause of dying because a statement on the causes of death refers to people who have lived. A less evident example is the syphilis example of Scriven (1959). *Treponema pallidum*, a bacteria, is the unique cause of syphilis. However, only a small percentage of people contaminated by the syphilis bacteria come into the third and last phase of *paralysis generalis*, a brain paralysis accompanied by motoric disorder and mental disturbances and causing death. Now, the first statement about the unique cause refers to a different causal field than the second statement about *paralysis generalis*. The first is the causal field of all persons who are susceptible to the bacteria, for example, all persons who have sexual intercourse. The second is the causal field of all persons who are already contaminated by the bacteria.

In research practice, this notion of causal field is of crucial importance because it also contains, next to self-evident fixed factors related to the research problem, factors that are fixed due to pragmatic considerations related to time and space (living in a country and doing research in that country) in addition to factors that are actually INUS conditions but based on causal a priori considerations and, due to the danger of heterogeneity, have been fixed (doing research in the causal field of the unemployed and leaving young workers out because youth unemployment is a different problem entirely). Capital sin number one in the social sciences is really the neglect of the causal field.

—Jacques Tacq

REFERENCES

- Aristotle. (1977). *Metaphysica*. Baarn: Het Wereldvenster.
- Bunge, M. (1959). *Causality: The place of the causal principle in modern science*. Cambridge, MA: Harvard University Press.
- Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. London: Collier-Macmillan.
- Hume, D. (1786). *Treatise on human nature*. Oxford, UK: Clarendon.
- Kant, I. (1781). *Kritik der reinen Vernunft* [Critique of pure reason]. Hamburg: Felix Meiner.
- Mackie, J. (1974). *The cement of the universe*. Oxford, UK: Oxford University Press.
- Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.
- Popper, K. (1972). *Objective knowledge: An evolutionary approach*. Oxford, UK: Clarendon.
- Russell, B. (1912). *Mysticism and logic*. London: Allen & Unwin.
- Scriven, M. (1959). Explanation and prediction in evolutionary theory. *Science*, 130, 477–482.
- Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North Holland.
- Tacq, J. J. A. (1984). *Causaliteit in Sociologisch Onderzoek. Een Beoordeling van Causale Analysetechnieken in het Licht van Wijsgerige Opvattingen over Causaliteit* [Causality in sociological research: An evaluation of causal techniques of analysis in the light of philosophical theories of causality]. Deventer, The Netherlands: Van Loghum Slaterus.
- Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, 5, 161–215.

CCA. See CANONICAL CORRELATION ANALYSIS

CEILING EFFECT

A ceiling effect occurs when a measure possesses a distinct upper limit for potential responses and a large concentration of participants score at or near this limit (the opposite of a FLOOR EFFECT). Scale attenuation is a methodological problem that occurs whenever variance is restricted in this manner. The problem is commonly discussed in the context of experimental research, but it could threaten the validity of any type of research in which an outcome measure demonstrates almost no variation at the upper end of its potential range.

Ceiling effects can be caused by a variety of factors. If the outcome measure involves a performance task with an upper limit (such as number of correct responses), participants may find the task too easy and score nearly perfect on the measure. Rating scales also possess upper limits, and participants' responses may fall outside the upper range of the response scale.

Due to the lack of variance, ceiling effects pose a serious threat to the validity of both experimental and nonexperimental studies, and any study containing a ceiling effect should be interpreted with caution. If a dependent variable suffers from a ceiling effect, the experimental or quasi-experimental manipulation will appear to have no effect. For example, a ceiling effect may occur with a measure of attitudes in which a high score indicates a favorable attitude and the highest response fails to capture the most positive evaluation possible. Because the measure cannot assess more favorable attitudes, a persuasion study using this measure may yield null findings if the prior attitudes of the participants were already at the high end of the attitude distribution.

Ceiling effects are especially troublesome when interpreting INTERACTION EFFECTS. Many interaction effects involve failure to find a significant difference at one level of an independent variable and a significant difference at another. Failure to find a significant difference may be due to a ceiling effect, and such interactions should be interpreted with caution.

Ceiling effects are also a threat for nonexperimental or correlational research. Most inferential statistics are based on the assumption of a normal distribution of the variables involved. It may not be possible to compute a significance test, or the restricted range of variance may increase the likelihood of rejecting the null hypotheses when it is actually true (TYPE II ERROR).

The best solution to the problem of ceiling effects is pilot testing, which allows the problem to be identified early. If a ceiling effect is found, the problem may be addressed several ways. If the outcome measure is task performance, the task can be made more difficult to increase the range of potential responses. For rating scales, the number of potential responses can be expanded. An additional solution involves the use of extreme anchor stimuli in rating scales. Prior to assessing the variable of interest, participants can be asked to rate a stimulus that will require them to use the extreme ends of the distribution. This will encourage participants to use the middle range of the rating scale, decreasing the likelihood of a ceiling effect.

—Robert M. Hessling,
Nicole M. Traxel, and Tara J. Schmidt

CELL

The term *cell* is used to refer to the point at which, in any CONTINGENCY TABLE, a row and a column intersect. Thus, a contingency table relating to two binary VARIABLES will have four cells.

—Alan Bryman

CENSORED DATA

Statistical data files are described as censored when four conditions are met: (a) data are available for a PROBABILITY SAMPLE of a POPULATION, (b) the DEPENDENT VARIABLE has one or more limit values, (c) the values of the INDEPENDENT VARIABLES are available for all cases regardless of the value of the dependent variable for any case, and (d) the limit value is not due to the selection of this value by the cases. A limit value is a value of a dependent variable either below or above which no other values can be measured or observed. As an example of a lower limit, neighborhood crime rates cannot be less than zero. Among neighborhoods with crime rates of zero, there could be two types of areas that have substantial differences in "crime proneness." Neighborhoods of the first type could be so safe that they would virtually never have a crime occur; those of the second type could have several "crime-facilitating"

characteristics but, just by chance alone, did not have a crime in a particular year. Despite the differences between these areas, both types would have a crime rate of zero as their lower limit, and there would not be any value of the dependent variable that distinguishes these neighborhoods from each other. Sellouts of various types of events are examples of upper limits on sales as a measure of total demand. Some events may have more unmet demand than others, but there is no way of telling.

Breen (1996) describes dependent variables with limit values as truncated. If cases have been not been excluded from the sample being studied because of their values on the dependent or independent variables, then the data as a whole meet the third condition for being described as censored in the sense that measurement of the dependent variable either below a lower limit or above an upper limit is not possible and is, therefore, censored. For neighborhood crime, the data are censored if all neighborhoods in a city are included in a sample regardless of their crime rate and regardless of the values of the independent variables. If a study excludes neighborhoods based on the value of a variable, then the data are truncated, not censored. The data for a study of only high-crime or low-crime neighborhoods, or only minority or nonminority areas, would be described as truncated because data collection was stopped or truncated at a certain value of a variable or characteristic of the neighborhoods.

To meet the fourth condition for being censored, the cases with limit values must not be the result of any self-selection by the cases themselves, as can happen with data on individuals. As Breen (1996) noted, households who do not buy luxury items have chosen or selected themselves not to be buyers, whereas neighborhoods do not decide whether they will have crime in them, regardless of how many security precautions residents may take. For the buying example, the analyses of the purchasing data can require sample selection rather than censored data methods.

—Dennis W. Roncek

See also CENSORING AND TRUNCATION

REFERENCE

Breen, R. (1996). *Regression models: Censored, sample selected, or truncated data*. Thousand Oaks, CA: Sage.

CENSORING AND TRUNCATION

Both *censoring* and *truncation* refer to situations in which empirical data on random variables are incomplete or partial. These terms have not been consistently used throughout the existing literature, and what one author calls *censoring*, another may term *truncation*. Consequently, an author's definition of one of these terms always requires careful attention.

Usually, censoring refers to a situation in which some values of a random variable are recorded as lying within a certain range and are not measured exactly, for at least some members of a sample. In contrast, truncation customarily refers to a situation in which information is not recorded for some members of the population when a random variable's values are within a certain range. Thus, censoring is associated with having incomplete or inexact measurements, whereas truncation is associated with having an incomplete sample, that is, a sample chosen conditional on values of the random variable. Hald (1952) is one of the first authors to distinguish between censoring and truncation. Since then, statistical methods for analyzing censored variables and truncated samples have been developed extensively (Schneider, 1986).

A censored *variable* is one in which some observations in a sample are measured inexactly because some values of the variable lie in a certain range or ranges. A censored *observation or case* refers to a sample member for which a particular variable is censored (i.e., known only to lie in a certain range).

For example, everyone whose actual age was reported as 90 years or older in the 1990 U.S. Census of Population has been censored at age 90 in the census data made publicly available. The U.S. Census Bureau decided to censor age at 90 to ensure confidentiality and to avoid inaccurate information (because very old people sometimes exaggerate their age). Any person whose actual age was reported to be 90 to 120 on the household census form appears in the public micro-level data as a case with age censored at 90.

A truncated *variable* is one in which observations are measured only for sample members if the variable's value lies in a certain range or ranges; equivalently, it is a variable for which there are no measurements at all if the variable's value does not fall within selected ranges. Conceptually, a truncated sample is closely related to a sample that has been trimmed to eliminate extreme values that are suspected of being outliers.

One difference between truncation and trimming is that samples are customarily truncated on the basis of a variable's numerical values but trimmed on the basis of a variable's order statistics (e.g., the top and bottom five percentiles are excluded, even though they exist in the data). For example, very short people are typically excluded from military service. As a result, the height of a sample of military personnel is a truncated variable because people in the entire population are systematically excluded from the military if their height is below a certain level.

TYPES OF CENSORING

There are different ways of classifying censoring and censored variables. One basic distinction is between random censoring and censoring on the basis of some predetermined criteria. The two common predetermined criteria used to censor variables are called Type I and Type II censoring. (A few authors also refer to Types III and IV censoring, but the definitions of other types are less consistent from author to author.)

Censoring that results because a variable's value falls above or below some exogenously determined value is called length-based or Type I censoring. When the U.S. Census Bureau censored people's ages at 90 years old, as described above, it performed length-based censoring.

Type II censoring occurs when sample values of the variable are ordered (usually from the lowest to the highest value), and exact measurements cease after the ordered sample values have been recorded for a predetermined fraction of the sample. Type II censoring mainly occurs when the variable refers to time (e.g., time to an event, such as death), and measurements terminate after the time of the event has been recorded for a preselected fraction of the sample. The remaining sample members are known only to have times exceeding the time at which the measurement process stopped (i.e., greater than the last observed time). For example, social scientists might record the duration of joblessness for 1,000 people who have just lost their jobs because a factory closed. If the researchers decide a priori to stop collecting data on the duration of joblessness after 60% of the sample members have found new jobs, they are implementing Type II censoring based on the duration of joblessness. The fraction of censored cases is fixed in advance (40% in this example); however, the maximum observed duration of joblessness depends on

the length of time that it takes 600 of the 1,000 people in the sample to find new jobs.

Often, censoring is the result of some other random process. To avoid biased conclusions, the random variable of interest must be statistically independent of the random process causing censoring. If the two are statistically independent, censoring is said to be noninformative; if they are statistically dependent, censoring is said to be informative.

Suppose, for example, that the variable of interest is a woman's age at the birth of her first child. Some women die childless; for these women, age at first birth is censored at their age at death. Death is clearly a random process. However, a key issue is whether a woman's age at death is statistically independent of her age at the birth of her first child. There is no statistical basis for determining if the process generating the outcome of interest and the random censoring process are statistically independent. Consequently, decisions about their independence are made in practice on the basis of substantive arguments. In any given application, some scholars may argue that the two processes are statistically independent, whereas others may argue that they are dependent. Still others may argue that these two processes are statistically independent (for all practical purposes) under some conditions but are dependent under other conditions. For example, the process governing the birth of a woman's first child and the process causing women to die may be (almost) independent in a developed country such as Sweden but not in a less developed country in sub-Saharan Africa.

Right, left, double, and multiple censoring are also ways of distinguishing censoring patterns. Right censoring means that values of the random variable are not observed exactly if they exceed a certain level τ_R ; left censoring means that they are not observed exactly if they are less than a certain level τ_L . Double censoring refers to the combination of left and right censoring; it means that values are observed when they are above a certain level τ_L and below some other level τ_R , that is, between τ_L and τ_R . Finally, multiple censoring occurs when there are multiple censoring levels, and the data record only the intervals in which the values of the variable fall.

For example, a person's annual earnings may be right censored at \$100,000 (i.e., higher earnings are recorded as censored at \$100,000) or left censored at \$5,000 (i.e., lower earnings are recorded as censored at \$5,000). In addition, earnings may be recorded exactly for values between \$5,000 and \$100,000 but

censored at these two extremes; then there is double censoring. Finally, surveys often ask people to report their annual earnings in terms of several ordered ranges (e.g., less than \$5,000, \$5,000–\$10,000, \$10,000–\$20,000, . . . , \$90,000–\$100,000, more than \$100,000). Then the responses are multiply censored; an ordinal variable has been recorded instead of the original continuous variable.

ANALYSES OF CENSORED OUTCOMES

Tobin's (1958) suggestion of a censored normal regression model of consumer expenditures was among the first to incorporate the idea of censoring into empirical social scientific research. In the censored normal regression model, the true dependent variable Y^* is fully observed above some level τ_L (and/or below some other level τ_R); otherwise, it is known only to be below τ_L (or above τ_R , or outside the range from τ_L to τ_R). That is, the data on the continuous variable Y may be left censored, right censored, or censored on the left and the right. Formally, these three cases mean the following.

1. Censored on the left:

$$\begin{aligned} Y &= Y^*, & \text{if } Y^* \geq \tau_L \\ &= \tau_L, & \text{if } Y^* < \tau_L. \end{aligned}$$

2. Censored on the right:

$$\begin{aligned} Y &= Y^*, & \text{if } Y^* \leq \tau_R \\ &= \tau_R, & \text{if } Y^* > \tau_R. \end{aligned}$$

3. Censored on the left and the right:

$$\begin{aligned} Y &= Y^*, & \text{if } \tau_L \leq Y^* \leq \tau_R \\ &= \tau_L, & \text{if } Y^* < \tau_L \\ &= \tau_R, & \text{if } Y^* > \tau_R. \end{aligned}$$

It is assumed that Y^* has a cumulative probability distribution function, $F(Y^*)$. If Y^* has a normal distribution and the mean of $Y^* = X'\beta$, then the model is called a TOBIT model. This model can easily be estimated by the method of maximum likelihood. Other censored regression models are also readily estimated by maximum likelihood as long as the cumulative probability distribution of Y^* can be specified.

If the true dependent variable Y^* is subject to multiple, interval censoring at n points, $\tau_1 < \tau_2 <$

$\tau_3 < \dots < \tau_n$, and once again $F(Y^*)$ has some known or postulated distribution, then REGRESSION MODELS FOR ORDINAL VARIABLES can also be estimated by maximum likelihood. If $F(Y^*)$ is the cumulative normal distribution, the model is known as an ordinal probit model; if $F(Y^*)$ is the cumulative logistic distribution, the model is known as the ordinal logit model (Maddala, 1983).

In the same year that Tobin proposed the censored normal regression model, Kaplan and Meier (1958) proposed a product-limit estimator of $(1 - F(t))$ for right-censored data on the length of the random time T to an event (e.g., time until death) when the probability distribution of T , $F(t)$, was not specified a priori. Their proposed estimator, now known as the Kaplan-Meier estimator, which is unbiased and asymptotically normal, triggered a huge growth in treatments of censoring in biostatistical, engineering, and social scientific analyses of censored data on the time to an event, such as death or failure.

TRUNCATION

Similarly to censoring, samples may be truncated on the left, right, or doubly. A sample is truncated on the left when members of the population are excluded from a sample for values of the variable less than τ_L . A sample is truncated on the right when members of the population are excluded for values of the variable greater than τ_R . Double truncation occurs when members of the population are excluded if the values of the variable are less than τ_L or greater than τ_R (i.e., the values are not between τ_L and τ_R). In some instances, τ_L and τ_R are known a priori; in other instances, τ_L and τ_R are unknown and are estimated from the sample data, along with other sample statistics for the random variable.

Analyzing data from a truncated sample is almost always more problematic than analyzing censored data. When data are censored, one knows the fractions of the sample that have complete and incomplete measurements, as well as the value(s) at which censoring occurs. When analyzing data from a truncated sample, there are usually lingering doubts about the true distribution of the variable in the entire population. Furthermore, one rarely knows the fraction of the population that is excluded because of their values on the variable of interest. However, if the true probability distribution of the variable is known, one can estimate a model of the conditional distribution of the outcome,

given that it is greater than τ_L , less than τ_R , or between τ_L and τ_R , depending on whether there is left, right, or double truncation.

Wachter and Trussell (1982) discussed estimating the distribution of heights of men in 18th-century England using historical data on recruits to the British navy, who were required to be taller than a certain level. This requirement meant that the naval recruits were a truncated sample of 18th-century Englishmen. In their application, they assumed that height has a normal distribution because a normal distribution fits data on people's heights in other, nontruncated samples very well. They dealt with other unusual problems, such as random fluctuations in the truncation level that was used, heaping at certain values, and rounding.

—Nancy Brandon Tuma

REFERENCES

- Hald, A. (1952). *Statistical theory with engineering applications*. New York: John Wiley.
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.
- Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.
- Schneider, H. (1986). *Truncated and censored samples from normal populations*. New York: Marcel Dekker.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24–36.
- Wachter, K. W., & Trussell, J. (1982). Estimating historical heights. *Journal of the American Statistical Association*, 77(378), 279–293.

CENSUS

A census is an exercise in which an entire statistical POPULATION is counted. Although the term can be applied to many types of population (e.g., a traffic census counts all vehicles passing a particular point over a particular period), it is usually used, and is used here, to refer to a census of population—a demographic exercise that involves the counting of the entire population of a specified geographical area, usually a country, over a short period of time. Censuses are both CROSS-SECTIONAL (they provide results for a single point in time) and LONGITUDINAL (they provide a series of regular counts that allow the characteristics of places to be tracked over time).

Censuses are the most ubiquitous social research exercises carried out around the world. The United Nations coordinates census activities by disseminating good practice guidelines and definitions and publishing comparable population counts for most countries of the world.

The minimum information collected in a census relates to the age, sex, and household composition of the population. Further personal characteristics that are collected typically include ethnicity, employment status, and educational qualifications. In addition, many countries also collect information about housing characteristics. Some countries include other issues such as religious denomination or details of journeys to work. Financial information, particularly income, is not consistently collected around the world because of reluctance, in certain cultures, to reveal such information.

The organizational challenges of a census are formidable and involve organizing a field workforce large enough, sufficiently well trained, and sufficiently well coordinated to visit every location in a country where people may have stayed on census day or, more usually, over the night of the census. Such a workforce needs to be organized using a geographical framework that ensures that every dwelling space (which, in addition to houses, communal buildings, and hotels, must include areas where the homeless sleep on the streets or without shelter or where people work over census night) is included.

To collect census data effectively, one must have a complete list of the locations where information needs to be collected. Every dwelling and those resident at each dwelling must only be counted once or, if they are counted twice (e.g., students may be counted at both their home and their university locations), that must be noted and appropriate adjustments must be made in the total population count. This is achieved by having a geographical framework for a census.

The most frequently used framework is based on a map, preferably of large enough scale to show every potentially occupied building. In countries without adequate mapping, high-resolution satellite or aerial photography can be used. The map is divided into a number of nonoverlapping districts, each of which will be visited by a single census enumerator or a single enumeration team. These areas must completely fill the map, ensuring that every location that needs to be covered is included and that every place to be visited is allotted only once. These “enumeration districts”

constitute the “input” geography of the census. Census results are also often published for enumeration districts. Modern censuses also record an accurate map reference (latitude and longitude, or local coordinates) for every house or building visited. These can either be read from the map or recorded using a Global Positioning System (GPS) receiver.

The advantage of collecting such map references, known as geo-codes, for every census location is that they allow census figures to be calculated simply (usually using a geographical information system) for any desired “output” geography. This allows the areas for which census data are published to be different from the “input” units, which need to be defined for the efficient management of the data collection teams rather than for meeting the needs of census data users.

Alternatives to systems of census geography that are based on maps are systems based on address lists and associated street maps. These require the production of a reliable “master address file” prior to the census. This needs to be complete and easy to relate to a map. A good example of this approach is the TIGER (Topologically Integrated Geographic Encoding and Referencing) system used in the United States, which is now being augmented with the Master Address file (MAF), which allows census forms to be mailed out rather than delivered by enumerators. A difficulty with address files is keeping them sufficiently up-to-date to provide a reliable base for the census. The enumeration process validates and completes such files.

Censuses need to be as accurate as possible because the results are used for an important set of legal, administrative, and research purposes. In the United States, the obligation to carry out a simple head count census is enshrined in the Constitution so that the boundaries of the areas that elect a member of Congress can be redrawn. This ensures fair elections based on equal populations electing each representative. Census results are used to allocate central government funds to local administrations based on the size of their population and the needs of that population. Censuses also provide the quantitative base for a very large number of other social research exercises by providing a “denominator.” This is the figure for the base population of an area so that rates can be calculated. For example, a count of how many unemployed people are claiming benefits in an area is not very meaningful unless one can estimate what percentage of the total economically active population in that area the count represents. The census provides that base figure, allowing percentage

unemployment rates to be calculated and compared reliably between areas.

Although censuses are an internationally comparable way of achieving population estimates, they are becoming more difficult to carry out, particularly in urban areas with a mobile and hard-to-count population. Censuses intrude on the PRIVACY of those filling in census forms; most citizens accept this intrusion as a necessary part of efficient government or because there may be penalties for noncompliance, but they are becoming more difficult to conduct. Census agencies apply very high standards of confidentiality to census data to satisfy the population that the invasion of their privacy will not be abused; nevertheless, it is becoming increasingly difficult to count certain parts of the population. For example, immigrants whose legal status is questionable are unlikely to willingly be counted; it also appears to be the case that in some countries, single young men are difficult to count because they are unlikely to be at home when enumerators call and have little incentive to fill in census forms.

Because of the increasing costs and difficulties in conducting censuses, many countries are investigating and implementing alternative register-based systems of tracking populations. Such registers may either be specific population registers that require citizens to register any change of address or family composition, or they may be based on other routinely collected information such as health registration, voter registration, property tax details, or drivers’ licenses. It appears likely that register-based systems will become the norm in most developed countries in the near future, with censuses, or “sample censuses” (more accurately, “population surveys” based on drawing a statistical sample of the population), being used to ensure that register-based data are accurate. This will have the advantage of keeping better track of populations over time than 10 yearly, or even 5 yearly, censuses can achieve.

The census remains the “gold standard” of demographic research in most countries and is usually the single most important source of social data. However, the move to a “knowledge-based society,” in which information technology is all-pervasive and information about individuals from many sources can be combined reliably and legally, will eventually make the census obsolete. Until then, it is the basis of demographic science and provides the base data on which much other social research is based.

—Robert Barr

REFERENCES

- Barr, R. (1996). A comparison of aspects of the US and UK censuses of population. *Transactions in GIS*, 1, 49–60.
- Dale, A., & Marsh, C. (Eds.). (1993). *The 1991 census user's guide*. London: HMSO.
- National Research Council. (2001). *The 2000 census: Interim assessment*. Washington, DC: National Academy Press.
- Rees, P., Martin, D., & Williamson, P. (Eds.). (2002). *The census data system*. Chichester, UK: Wiley.
- United Nations Statistics Division. (2001). *Handbook on census management for population and housing censuses—Series F* (No. 83, Revision 1). New York: United Nations.

CENSUS ADJUSTMENT

The U.S. census tries to enumerate all residents of the United States, block by block, every 10 years. (A block is the smallest unit of census geography; the area of blocks varies with POPULATION density. There are about 7 million blocks in the United States.) State and substate counts matter for apportioning the House of Representatives, allocating federal funds, congressional redistricting, urban planning, and so forth. Counting the population is difficult, and two kinds of ERROR occur: *gross omissions* (GOs) and *erroneous enumerations* (EEs). A GO results from failing to count a person; an EE results from counting a person in error. Counting a person in the wrong block creates both a GO and an EE. Generally, GOs slightly exceed EEs, producing an undercount that is uneven demographically and geographically. In 1980, 1990, and 2000, the U.S. Census Bureau tried unsuccessfully to adjust census counts to reduce differential undercount using *dual-systems estimation* (DSE), a method based on CAPTURE-RECAPTURE. (Some other countries adjust their censuses using different methods.) For discussion, see the special issues of *Survey Methodology* (1992), *Journal of the American Statistical Association* (1993), *Statistical Science* (1994), and *Society* (2001). DSE involves the following:

- taking a random sample of blocks (in 2000, DSE sampled 25,000 blocks);
- trying to enumerate the residents of those blocks after census day, in the Post-Enumeration Survey or PES;
- trying to match PES records to census records, on the basis of data that often are incomplete or erroneous;

- estimating the undercount within demographic groups called *post strata* (in 2000, DSE used 448 post strata) by comparing capture-recapture estimates with census counts.

Being counted in the census is considered capture; being counted in the PES is considered recapture. The PES attempts to identify EEs and to account for movers. Ultimately, this yields an adjustment factor for each post stratum: the estimated population of the post stratum divided by the census count for the post stratum. The adjustment for each block in the country is found by separating the census count in that block into its components, post stratum by post stratum; applying the adjustment factor for each post stratum to the corresponding count; and then summing the adjusted counts to get the adjusted total for the block. This procedure is justified by the synthetic ASSUMPTION that the undercount rate is constant within each post stratum, regardless of geography. Failure of the synthetic assumption is called HETEROGENEITY.

There is another way to estimate the population, called *demographic analysis* (DA). DA estimates the population from administrative records and estimates of immigration and emigration using the identity

$$\begin{aligned} \text{population} &= \text{births} - \text{deaths} \\ &+ \text{immigration} - \text{emigration}. \end{aligned}$$

Historically, DA estimates were the primary evidence of net census undercount and of differential undercount by race—motivating census adjustment.

In 1980, the U.S. Census Bureau decided not to adjust the census: Too many data were missing. In 1990, the Bureau sought to adjust, but the administration overruled the Bureau, finding that adjustment was unlikely to improve accuracy. The Bureau planned to adjust the 2000 census but in the end decided not to because DSE disagreed with DA. According to the 2000 census, there were about 281.4 million people residing in the United States. Adjustment would have added 1.2% to this number, but DA indicated that the census had found 0.7% too many people.

The U.S. census is remarkably accurate; proposed adjustments are relatively small. For example, in 2000, adjustment would have increased the population share of Texas by 0.043%, from 7.4094% to 7.4524%. That would have been the biggest state share change. Adjustment must be extremely accurate for such tiny changes to improve the census. Response errors in the census or the PES lead to problems in matching records and

produce *processing errors* in DSE that are large on the relevant scale. Heterogeneity and CORRELATION bias also produce large errors.

DSE begins with capture-recapture but has layer upon layer of complexity in which details have big effects on the population estimate. For example, there are procedures for getting data by proxy interviews, searching neighboring blocks for missing records, detecting duplicate records, accounting for people who move between census day and the PES, and imputing missing data. The keys to DSE are matching PES records to census records accurately, independence, and the synthetic assumption—not counting better.

DSE would have added 3.3 million people net to the 2000 census. As of October 2001, the Census Bureau estimated net processing error in DSE to be 5 to 6 million and gross error in the DSE to be more than 12 million. In comparison, gross census error was estimated to be about 10 million (U.S. Census Bureau, 2001). Error in the adjustment is at least as large as the census error that DSE is intended to fix: Adjusting the U.S. census could easily make it worse.

—Philip B. Stark

REFERENCES

- Freedman, D. A., & Wachter, K. W. (2001). *On the likelihood of improving the accuracy of the census through statistical adjustment* (Tech. Rep. 612). Berkeley: University of California, Department of Statistics.
- The 1980 U.S. Census [Special issue]. (1992). *Survey Methodology*, 18.
- The 1990 U.S. Census [Special issue]. (1993). *Journal of the American Statistical Association*, 88.
- The 1990 U.S. Census [Special issue]. (1994). *Statistical Science*, 9.
- The 2000 U.S. Census. (2001). *Society*, 39, 2–53.
- U.S. Census Bureau. (2001). *Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy on Adjustment for Non-Redistricting Uses* (With supporting documentation, Reps. 1–24). Washington, DC: Government Printing Office.

CENTRAL LIMIT THEOREM

This is a theorem that shows that as the size of a sample, n , increases, the SAMPLING DISTRIBUTION of a statistic approximates a NORMAL DISTRIBUTION even when the distribution of the values in the population are skewed or in other ways not normal. In addition, the

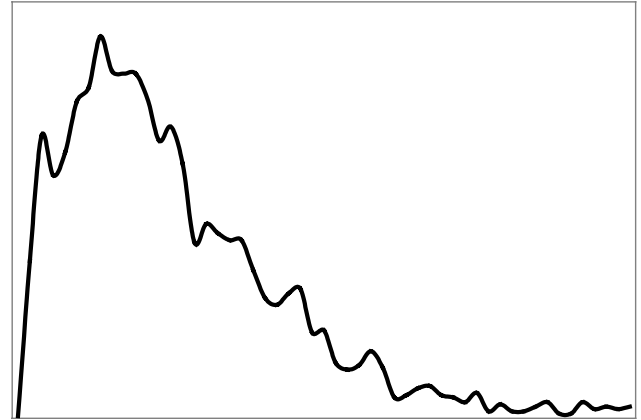


Figure 1 An Example of a Skewed Distribution

spread of the sampling distribution becomes narrower as n increases; that is, the bell shape of the normal distribution becomes narrower. It is used in particular in relation to the sample mean: Thus, the sampling distribution of a sample mean approximates a normal distribution centered on the true population mean. With larger samples, that normal distribution will be more closely focused around the population mean and will have a smaller standard deviation. The theorem is critical for statistical inference as it means that population parameters can be estimated with varying degrees of confidence from sample statistics. For if the means of repeated samples from a population with population mean μ would form a normal distribution about μ , then for any given sample, the probability of the mean falling within a certain range of the population mean can be estimated. It does not require repeat sampling actually to be carried out. The knowledge that repeated sampling produces a normal and (for larger n) narrow distribution of sample averages concentrated around the population mean enables us to predict at varying levels of confidence just how close to the true mean our sample mean is. For large samples, the relationship will hold even if the population distribution is very skewed or discrete. This is important because in the social sciences, the distributions of many variables commonly investigated (e.g., income) are skewed. Large n can mean as little as 30 cases for estimation of a mean but might need to be more than 500 for other statistics.

To illustrate, Figure 1 shows an example of a skewed distribution. It is a truncated British income distribution from a national survey of the late 1990s. The means of around 60 repeat random samples of sample size

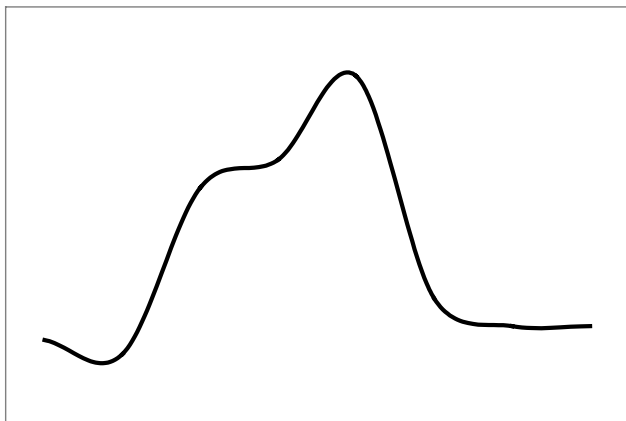


Figure 2 The Distribution of Sample Means From the Distribution Shown in Figure 1, in Which Sample Size = 10

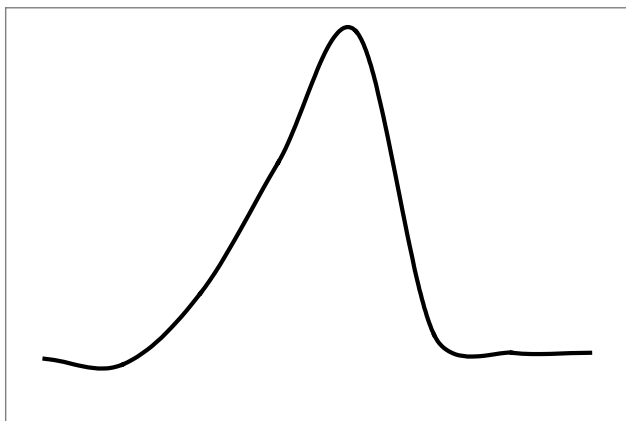


Figure 3 The Distribution of Sample Means From the Distribution Shown in Figure 1, in Which Sample Size = 100

$n = 10$ were taken from this distribution and result in the distribution shown in Figure 2. This distribution is less skewed than that in Figure 1, though not exactly normal. If a further 60 repeat random samples—this time, of $n = 100$ —are taken from the distribution illustrated in Figure 1, the means of these samples form the distribution in Figure 3, which clearly begins to approximate a normal distribution.

EXAMPLE

As a result of this theorem, we can, for example, estimate the population mean of, say, the earnings of female part-time British employees. A random sample

of 179 such women in the late 1980s provided a sample mean of earnings of £2.57 per hour with a sample standard deviation, SD , of 1.24, for the particular population under consideration. We can take the sample standard deviation to estimate the population standard deviation, which gives us a standard error of $1.24/\sqrt{179} = 0.093$. On the basis of the calculated probabilities of a normal distribution, we know that the population mean will fall within the range $2.57 \pm (1 \times 0.093)$, that is, between £2.48 and £2.66 with 68% probability; that it falls within the range $2.57 \pm (1.96 \times 0.093)$, that is, between £2.39 and £2.75 with 95% probability; and that it falls within the range $2.57 \pm (2.58 \times 0.093)$, that is, between £2.33 and £2.81 with 99% probability.

We can be virtually certain that the range covered by the sample mean, plus or minus 3 standard errors, will contain the value of the population mean. But as is clear, there is a trade-off between precision and confidence. Thus, we might be content being 95% confident that the true mean falls between £2.39 and £2.75 while acknowledging that there is a 1 in 20 chance that our sample mean does not bear this relationship to the true mean. That is, it may be one of the samples whose mean falls into one of the tails of the bell-shaped distribution.

HISTORICAL DEVELOPMENT

The central limit theorem can be traced to the year 1810 and to a proof offered by mathematician Pierre Laplace, although it was, itself, a development of Abraham De Moivre's limit theorem. The appeal of the theorem was immediate and constituted a significant step in statistical understanding. The binomial and the Poisson distributions are instances of the theorem in operation. However, a caveat was presented by the Cauchy density, named after Augustin Cauchy but also identified by Poisson. This is an instance of a distribution, with an undefined mean and standard deviation, for which the central limit theorem does not hold. The existence of this distribution indicates the existence of exceptions to the large sample rule and the fact that the theorem is based on certain assumptions about the underlying distribution, which can be shown not to be fulfilled in every case.

Nevertheless, the central limit theorem remains critical to statistical inference and estimation in that it enables us, with large samples, to make inferences based on assumptions of normality. The theorem can

also be proved for multivariate variables under similar conditions as those that apply to the univariate case.

—Lucinda Platt

See also BELL-SHAPED CURVE, NORMAL DISTRIBUTION, SAMPLING DISTRIBUTION

REFERENCES

- Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.
- Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics: Vol. 1. Distribution theory* (5th ed.). London: Griffin.

CENTRAL TENDENCY. See MEASURES OF CENTRAL TENDENCY

CENTROID METHOD

The *centroid method* is an agglomerative CLUSTERING method, in which the similarities (or dissimilarities) among clusters are defined in terms of the *centroids* (i.e., the multidimensional means) of the clusters on the variables being used in the clustering. Specifically, the centroid method is a sequential agglomerative hierarchical clustering or SAHN ALGORITHM (Sneath & Sokal, 1973), that is, a method in which all entities to be clustered are initially separate, then are combined in steps to form higher order clusters. The centroid method was first described by Sokal and Michener (1958).

The input for the centroid method is a set of N entities or “cases” to be clustered, in which each case (denoted O_i) is a multivariate observation on a set of relevant numeric descriptive variables, $X_1, \dots, X_M$. In other words,

$$O_i = [X_{i,1}, X_{i,2}, \dots, X_{i,M}].$$

The output of the method is an assignment of cases to a hierarchical (and nested) set of clusters. Usually, this hierarchical clustering solution is represented by a TREE DIAGRAM or “dendrogram.”

DESCRIPTION OF THE ALGORITHM

As in any agglomerative clustering method, the centroid method algorithm proceeds in iterative steps. At the first step, there are N entities or cases to be clustered. However, at each step, two similar cases are combined, resulting in a cluster that is itself then treated as a new case (and the constituent cases are removed). Thus, the next step of the algorithm proceeds on the reduced set of $N - 1$ cases.

At the first step, the algorithm begins by defining the dissimilarity between each pair of cases O_i and O_j as the Euclidean distance (or squared Euclidean distance) between the two cases:

$$d_{i,j} = \left[\sum_{m=1}^M (X_{i,m} - X_{j,m})^2 \right]^{\frac{1}{2}}.$$

Note that the values of case O_i on the variables $X_1, \dots, X_M$ can be considered to be the centroid of a trivial cluster with one member. The result of these distance calculations is a matrix **D** of dissimilarities between each pair of cases. The first step of the sequential agglomerative algorithm continues as follows. The pair of cases O_i, O_j with the smallest dissimilarity is selected to be combined, and these two cases are combined to form a cluster that can be considered to be a new case $O_{N'}$. This new case $O_{N'}$ is added to the set of cases being clustered, and its constituent cases O_i and O_j are removed. In other words, the multivariate observations O_i and O_j are replaced by a single new case $O_{N'}$, whose data values are the centroid of the cluster $O_{N'}$:

$$O_{N'} = [\bar{X}_{N',1}, \bar{X}_{N',2}, \dots, \bar{X}_{N',M}].$$

At this point, Step 1 of the iterative algorithm has been completed, and the next step commences. Note that the current number of cases is now $N - 1$, rather than N (because two cases have been replaced by a single new cluster, $O_{N'}$). For this reduced proximity matrix, the dissimilarity of the new cluster $O_{N'}$ to each other case O_k is defined as the Euclidean distance from the centroid of the cluster to the vector of values of case O_k on the variables. That is,

$$d_{N',k} = \left[\sum_{m=1}^M (\bar{X}_{N',m} - X_{k,m})^2 \right]^{\frac{1}{2}}.$$

The second step proceeds with the selection of the new closest pair of cases (i.e., the pair of cases with the

smallest dissimilarity) to be combined. At the second stage and later stages, either of the two cases being combined may be a cluster. The two closest cases are combined, the set of cases is reduced by one, and the proximity matrix is recomputed using the centroid of the new cluster.

These sequential steps are repeated until only two cases remain. These two cases may be combined into a single cluster containing all the cases.

WEIGHTED VERSUS UNWEIGHTED CENTROID METHOD

One complication has been omitted from the above description of the algorithm. This issue is how the centroid should be computed when the two cases being combined comprise different numbers of the original N cases. In the most common version of the algorithm, the centroid is computed using the data (equally weighted) from the original set of cases comprising that cluster. This version is referred to by Sneath and Sokal (1973) as the *unweighted pair-group centroid method*. On the other hand, if the centroid of a cluster is computed as the simple multidimensional average of the two cases being combined, without regard to the number of original cases comprising the two current clusters or cases, that has the effect of weighting individual cases in a larger cluster less than cases in a smaller cluster. This variant of the algorithm is termed the *weighted pair-group centroid method*.

An Example

A numerical example is presented in Table 1. In this hypothetical application, the data consist of math and verbal test scores for five students, and the Euclidean distance measure is used. At Step 1, the minimum distance ($= 82.46$) is between Cases $s03$ and $s04$. These two cases are averaged to form Cluster $c34$. At Step 2, the minimum distance between entities is between Case $s01$ and Cluster $c34$, so these are combined (using the unweighted method) to form Cluster $c134$ at Step 3. The algorithm proceeds in this manner until all cases are combined in a single cluster c^* .

NONMONOTONICITY

Monotonicity is a property that is satisfied by many sequential agglomerative hierarchical clustering methods, such as single-link or complete-link

Table 1

Step	Subject	Math	Verbal	Minimum Distance
1	$s01$	600	380	$d(s03, s04) = 82.46$
	$s02$	800	790	
	$s03$	760	330	
	$s04$	740	410	
	$s05$	520	780	
2	$s01$	600	380	$d(s01, c34) = 150.33$
	$s02$	800	790	
	$c34$	750	370	
	$s05$	520	780	
3	$c134$	700	373.33	$d(s02, s05) = 280.18$
	$s02$	800	790	
	$s05$	520	780	
4	$c134$	700	373.33	$d(c134, c25) = 413.61$
	$c25$	660	785	
5	c^*	684	538	

clustering. To define monotonicity, let $O_{N'}$ be a cluster formed by combining cases O_i and O_j . A hierarchical clustering method satisfies *monotonicity* if $d_{N',k}$, the dissimilarity of the new cluster $O_{N'}$ to any other case O_k , is always equal to or greater than the two dissimilarities $d_{N',i}$ and $d_{N',j}$. The centroid method, in both its unweighted and weighted variants, does not satisfy monotonicity (Anderberg, 1973; Milligan, 1979). The method has been criticized for this failure because satisfying monotonicity ensures that the tree diagram used to represent a hierarchical cluster structure can be plotted with the height of each node representing the dissimilarity between clusters at a particular level or stage of the algorithm.

APPLICABILITY IN THE SOCIAL SCIENCES

Because the input for the centroid method is a set of multivariate observations in which the variables are assumed to be numeric, it might be applicable when the data consist of a sample of patient records or subjects measured on a test with multiple subscales. Use of the Euclidean distance or squared Euclidean distance to define the dissimilarities between clusters implies that the variables or subscales ought to be measured on comparable scales. In the social sciences, the centroid method has been used less frequently than other SAHN techniques such as Ward's method.

—James E. Corter

REFERENCES

- Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.
- Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms. *Psychometrika*, 44(3), 343–346.
- Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification*. San Francisco: Freeman.
- Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409–1438.

CETERIS PARIBUS

Ceteris paribus is a Latin phrase meaning “other things being equal.” It therefore refers to the process of comparing like with like when asserting a causal relationship or the effect of one variable on another. For example, we might be interested in ascertaining whether certain forms of training increase subsequent earnings. To identify the specific effect of the training, however, we would need to CONTROL for a number of individual characteristics such as gender, qualifications, employment history, age, and so on, as well as perhaps local labor market conditions. We would then be assessing the effect of training on individuals who were matched on these other factors, that is, who had similar qualifications, similar work history, and so forth and only differed in whether they received additional training. If there remained a significant effect of training net of these controls, we could then say that, *ceteris paribus*, training increased earnings. In fact, when we use statistical methods to compare like with like, we do not need to have exactly matched individuals within our sample data to be able to claim effects, other things being equal. Rather, multivariate methods estimate the effect of each different factor net of each other. Then the effect of training will be that at a fixed position on all other variables (such as men compared with men at one of a particular level of qualifications, etc.), even if some of those combinations of fixed positions do not occur within our data.

For example, we might be interested in whether women have lower earnings than men. Table 1 shows the results from LINEAR REGRESSIONS on a sample of British data from the late 1980s in terms of hourly earnings. The question is, Do women earn less per hour than men, on average? As Model 1 in Table 1 shows,

Table 1 Multiple Regression Showing Effects on Average Hourly Wage of Sex, Occupation, and Education, Britain, Late 1980s

	Model 1	Model 2
Constant	4.97	3.94
Sex (Comparison = male)		
Female	−1.96***	−1.79***
Occupation (Comparison = manual occupations)		
Professional		2.23***
Intermediate		0.78**
Educational qualifications (Comparison = no qualifications)		
Vocational and intermediate		0.16 (ns)
Higher qualifications		0.93***

** $p < .01$; *** $p < .001$; ns = not statistically significant at the .05 level.

the answer would appear to be yes. They would appear to earn nearly £2 less per hour than men.

However, it might be argued that the reason for the lower earnings of women is due to the fact that they occupy different occupational niches than men and that their average levels of education also play a role, rather than the critical factor being sex per se. Therefore, to compare like with like, we control occupational level and educational level, which gives the results of Model 2 in Table 1. We now see that, *ceteris paribus*, the effect of sex is reduced to a difference of £1.79 per hour. On the other hand, occupying a professional occupation increases the earnings of men and women by £2.23 compared to being in a manual occupation. Some of the difference in hourly earnings between the sexes can therefore be attributed to different professional and educational patterns among the sexes. Only part, however. *Ceteris paribus*, we can still say that women are likely to earn substantially less than their male counterparts and that this difference is statistically significant.

—Lucinda Platt

See also CONTROL, MULTIVARIATE ANALYSIS

REFERENCE

- Lewis-Beck, M. S. (Ed.). (1993). *Regression analysis*. London: Sage/Toppan.

CHAID

CHAID (CHi-squared Automatic Interaction Detection) is a tree-based segmentation technique useful in the exploratory analysis of a categorical dependent variable Y in terms of a large number of categorical explanatory variables. The basic algorithm works by successively splitting the sample (the initial “parent” group) into smaller and smaller subgroups that differ significantly with respect to the dependent variable, thereby growing a tree. The process continues until one of the stopping rules is triggered.

At any given point in the analysis, a given parent group is evaluated for possible splitting into two or more subgroups according to that predictor determined to be the most statistically significant (lowest p value) in its relationship to the dependent variable. In the original algorithm (Kass, 1980), the p value for a predictor is computed using the chi-squared test for independence after merging those categories that do not differ

significantly with respect to the dependent variable and is adjusted using an appropriate Bonferroni multiplier.

Predictors are classified by the user as “free,” “monotonic,” or “floating” to establish those categories that are eligible to be merged; the merging algorithm combines categories, two at a time, in a stepwise fashion. A category of a free predictor is free to merge with any other categories of that predictor. For monotonic variables, only adjacent categories are eligible to be merged. Floating variables are treated as monotonic except for the final category, often representing missing data, which is eligible to merge with any other category.

Extensions include the following:

1. When Y is ordered categorical, the test for independence is replaced by the test for no Y association (Magidson, 1993). This test assesses the significance of differences between mean Y scores, analogous to ANOVA but without the normality assumption.

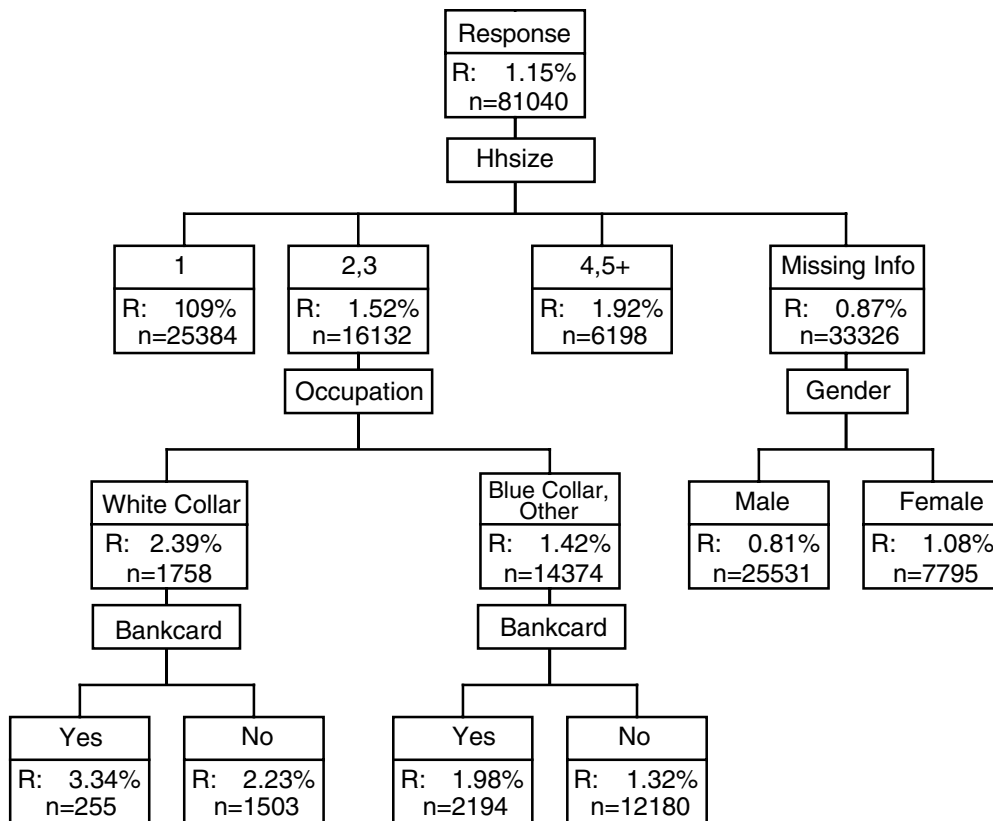


Figure 1 CHAID

2. For continuous dependent variables, the F statistic is used in place of the chi-squared statistic.
3. The merging algorithm is replaced by an exhaustive search through all possibilities (Biggs, De Ville, & Suen, 1991).
4. A hybrid CHAID adds the power of a model such as the LOGIT, or ordinal logit, to structure the tree, or latent class (LC) to extend the analysis to multiple dependent variables.

A latent class/CHAID hybrid analysis is used when there are multiple measures of response (profitability, purchase of different catalog items, quantity purchased), or in profiling the resulting segments obtained from a LATENT CLASS ANALYSIS. First, a LC analysis is performed to obtain an underlying latent variable that is passed on to CHAID as the dependent variable, which is then analyzed as a function of demographic or other exogenous variables.

Each node in the resulting tree predicts the latent class segments in addition to each of the original dependent variables, the latter obtained using LC model parameters. This type of analysis also results in a *common* splitting of the predictor categories appropriate for all the original dependent variables.

As an example of a CHAID/logit hybrid, data from Magidson (1993) was re-analyzed. The best segment using the original algorithm, HHSIZE = 2 or 3, OCCUP = "white collar," achieved a response rate of 2.39%. Figure 1 shows the result from a hybrid CHAID where the increased power of logit modeling allowed fitting a main-effect of CARD to both OCCUP nodes, resulting in an additional segment having a predicted response rate of 3.34%.

—Jay Magidson

REFERENCES

- Biggs, D., De Ville, B., & Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18, 49–62.
- Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2), 119–127.
- Magidson, J. (1993). The CHAID approach to segmentation modeling: CHi-squared Automatic Interaction Detection. In R. Bagozzi (Ed.), *Handbook of marketing research* (pp. 118–159). London: Blackwell.

CHANGE SCORES

A change score is the difference between the value of a variable measured at one point in time (Y_t) from the value of the variable for the same unit at a previous time point (Y_{t-1}). Change scores are also referred to as DIFFERENCE SCORES, or with the symbol ΔY , so that

$$\Delta Y = Y_t - Y_{t-1}. \quad (1)$$

Change scores are central to the analysis of PANEL DATA, or data collected more than once on the same individuals or units over time, with the goal of most analyses being to model how changes in Y are related to levels or changes in INDEPENDENT VARIABLES over time. Change scores are also the primary DEPENDENT VARIABLE in PRETEST/posttest EXPERIMENTAL DESIGNS.

One common approach to the analysis of change scores is to model ΔY as a function of the independent variables X_1 , X_2 , and so forth, as well as the value of Y itself at the earlier time period $t - 1$, as in

$$\Delta Y = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + \varepsilon, \quad (2)$$

where the β s represent REGRESSION COEFFICIENTS and ε the equation's ERROR term. The effects of X_{t-1} and other lag values of X may also be included in the MODEL. Y_{t-1} could be added to both sides of change score equation (2) to produce an equivalent model predicting the *level* of Y at time t :

$$Y_t = \beta_0 + \beta_1 X_t + (\beta_2 + 1)Y_{t-1} + \varepsilon. \quad (3)$$

Y_{t-1} , or the "LAGGED endogenous variable," is usually included in change score models for a variety of reasons. Perhaps the most important is due to the phenomenon in LONGITUDINAL data known as REGRESSION TOWARD THE MEAN. Regression effects reflect the tendency for large values of a variable at one point in time to be associated with greater subsequent change in the *negative* direction; as the large initial values were produced in part by large RANDOM ERRORS, they are likely to be smaller in the next time period. The inclusion of Y_{t-1} may also be justified on substantive grounds if it can be argued that prior values of Y are a CAUSE of its subsequent values.

Estimation of change score models is complicated by the presence of random MEASUREMENT ERROR in Y , in which case the observed change in Y will not accurately reflect the "true change" over time. Correcting for measurement error typically requires more

indicators of Y for two-wave data (i.e., additional variables that measure the same concept or LATENT VARIABLE as Y) or more waves of measurement for single-indicator models of Y (Kessler & Greenberg, 1981). Another potential problem is the presence of AUTOCORRELATED disturbances, in which the error terms for Y_t and Y_{t-1} are related, as occurs frequently in longitudinal data. In that case, Y_{t-1} will be related to the error term, and it will be necessary to estimate the model with an INSTRUMENTAL VARIABLE or related approach.

Change scores also figure prominently in models that attempt to control for FIXED EFFECTS, that is, characteristics of a given individual or unit that are unobserved by the researcher and that are stable over time. If Y_t is determined by a series of X s, unobserved stable variables Z , and an error term, then the potential BIASING effects of Z can be eliminated by constructing a change score model, that is, by subtracting the equation for Y_{t-1} from the equation for Y_t (or, in multiwave data, subtracting the Y for each unit in each year from that unit's grand mean). If Y_{t-1} is included as an independent variable in a dynamic fixed-effect model, however, it will necessarily be related to the model's error term, with instrumental variable estimation and multiwave data being necessary to correct this potential bias (Hsiao, 1986).

—Steven E. Finkel

See also DIFFERENCE SCORES

REFERENCES

- Hsiao, C. (1986). *Analysis of panel data*. Cambridge, UK: Cambridge University Press.
 Kessler, R., & Greenberg, D. (1981). *Linear panel analysis*. New York: Academic Press.

CHAOS THEORY

Chaos theory refers to a type of behavior with NON-LINEAR DYNAMICS that is both irregular and oscillatory. It is encountered mathematically with certain sets of nonlinear deterministic dynamic models in which patterns of overtime behavior are not repeated no matter how long the model continues to operate. Some discrete-time models using nonlinear difference equations are known to exhibit chaotic dynamics under certain conditions using only one equation. However,

in continuous-time models, the possibility of chaos normally requires a minimum of three independent variables, which usually requires an interdependent system of three differential equations. This requirement for continuous-time models can be changed in special cases if the system also has a forced oscillator or time lags, in which case chaos can appear even in single-equation continuous-time models.

Chaos can occur only in nonlinear situations. In multidimensional settings, this means that at least one term in one equation must be nonlinear while also involving several of the variables. Because most nonlinear models (and nearly all of the substantively interesting ones) have no analytical solutions, they must be investigated using numerically intensive methods that require computers. Because the dimensionality and nonlinearity requirements of chaos do not guarantee its appearance, chaotic behavior is typically discovered in a model through computational experimentation that involves finding variable ranges and parameter values that cause a model to display chaotic properties.

Chaotic processes may be quite common in real physical and even social systems, even though our present ability to identify and model such processes is still developing. Discovering real chaotic processes in physical and social systems is often quite difficult because STOCHASTIC noise is nearly always present as well in such systems, and it is not easy to separate truly RANDOM behavior from chaotic behavior. Nonetheless, mathematical tools continue to be developed that are aimed at sorting out these processes and issues.

Chaos has three fundamental characteristics. They are (a) irregular periodicity, (b) sensitivity to initial conditions, and (c) a lack of predictability. These characteristics interact within any one chaotic setting to produce highly complex nonlinear variable trajectories. Irregular periodicity refers to the absence of a repeated pattern in the oscillatory movements of the chaotically driven variables. Because of the irregular periodicity, Fourier analysis, graphing techniques, and other methods are commonly used to build a case for identifying chaotic processes (see Brown, 1995a).

The nonlinear model that has been among the most well studied with regard to chaos in discrete settings is a general form of a logistic map, and its chaotic properties were initially investigated by May (1976). This general logistic map is $Y_{t+1} = aY_t(1 - Y_t)$. Under

the right conditions, this map can produce the standard S-shaped trajectory that is characteristic of all logistic structures. However, oscillations in the trajectory occur when the value of the parameter a is sufficiently large. For example, when the value of parameter a is set to 2.8, the trajectory of the model oscillates around the equilibrium value of $Y_{t+1} = Y_t = Y^*$ while it converges asymptotically toward this equilibrium limit. But when the value of the parameter a is set equal to, say, 4.0, the resulting longitudinal trajectory never settles down toward the equilibrium limit and instead continues to oscillate irregularly around the equilibrium in what seems to be a random manner that is caused by a DETERMINISTIC process.

The most famous continuous-time model that exhibits chaotic behavior is the so-called “Lorenz attractor” (Lorenz, 1963). This model is an interdependent nonlinear system involving three first-order differential equations, and it was originally used to analyze meteorological phenomena.

Models with forced oscillators are sometimes good candidates for exhibiting chaotic or near chaotic (i.e., seemingly random) longitudinal properties. In the social sciences, such a model has been developed and explored by Brown (1995b, chap. 6). This model is a nonlinear system of four interdependent differential equations that uses a forced oscillator with respect to a parameter specifying alternating partisan control of the White House (or other relevant governmental institution). The dynamics of the system are investigated with regard to longitudinal damage to the environment, public concern for environmental damage, and the cost of cleaning up the environment. Variations in certain parameter values yield a variety of both stable and unstable nonlinear dynamic behaviors, including behaviors that have apparent random-like properties typically associated with chaos.

—Courtney Brown

REFERENCES

- Brown, C. (1995a). *Chaos and catastrophe theories*. Thousand Oaks, CA: Sage.
- Brown, C. (1995b). *Serpents in the sand: Essays on the nonlinear nature of politics and human destiny*. Ann Arbor: University of Michigan Press.
- Lorenz, E. N. (1963). Deterministic non-periodic flow. *Journal of Atmospheric Science*, 20, 130–141.
- May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, 26, 459–467.

CHI-SQUARE DISTRIBUTION

The chi-square distribution is a distribution that results from the sums of squared standard normal variables. Even though the sums of normally distributed variables have a NORMAL DISTRIBUTION (with the shape of a BELL-SHAPED CURVE), the sums of their squares do not. The commonly used chi-square distribution refers to one type known as the central chi-square distribution. The following discussion focuses on this type. As both negative and positive values square to a positive number, the curve only covers nonnegative numbers, with its left-hand tail starting from near zero and with a skew to the right. The curve is specified by its DEGREES OF FREEDOM (df), which is the number of unconstrained variables whose squares are being summed. The curve has its MEAN as the df , and the square root of twice the df forms its STANDARD DEVIATION. The highest point of the curve occurs at $df - 1$. As the number of df increases, the distribution takes on a more humpbacked shape with less skew but with a broader spread, given that its standard deviation will be increasing and that its furthest left point will remain at 0. Thus, a distribution with 5 df will have a mean of 5 and a standard deviation of 3.1, whereas a distribution with 10 df will have a mean of 10 and a standard deviation of 4.5. At high values of df , the curve tends toward a normal distribution, and in such (uncommon) cases, it is possible to revert to normal probabilities rather than calculating those for the chi-square distribution. The right tail probability of a chi-square distribution, which is the area in the right tail above a certain value, divided by the total area under the curve represents the likelihood that the chi-square value is randomly distributed. If the area in the tail beyond the given chi-square value represented, say, 5% of the area under the curve, then that chi-square value would be that at which we could be 95% confident that the association being tested held.

APPLICATION

These probabilities have been calculated and tabulated for certain ranges of chi-square values and df . Conversely, tables are commonly produced that provide the chi-square values that will produce certain probability levels at certain df . For example, such tables tell us that we can be 95% confident at a chi-square value of 9.49 for 4 df . Figure 1 shows the chi-square distribution for 4 df (with mean = 4 and standard

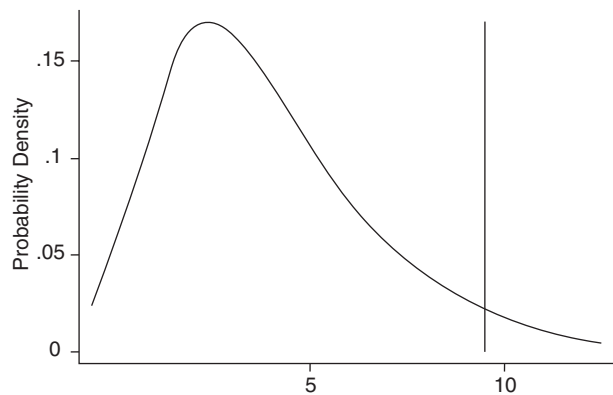


Figure 1 An Example of Chi-Square Distribution With 4 Degrees of Freedom

deviation = 2.83). The right tail probability is 5% at the point where the illustrated cutoff line takes the value of 9.49. This means that with statistics that approximate a chi-square distribution, we can compare the test statistic with such tables and ascertain its level of significance.

The most common such statistic is that resulting from Pearson's CHI-SQUARE TEST. However, this test is not the only use of the chi-square distribution. Other statistics that test the relationship between observed and "expected" or theoretical values (i.e., statistics that measure goodness of fit) approximate a chi-square distribution. For example, the likelihood ratio test statistic, also known as G^2 or L^2 , which is derived from maximum likelihood method, approximates a chi-square distribution. It does this through a (natural) log transformation of the ratio between the maximum likelihood based on the null hypothesis (the $-2 \log$ likelihood) and the maximum likelihood based on the actual data values found. G^2 is a statistic commonly quoted for log-linear models. In practice, chi-square and G^2 statistics, though distinct, tend to give similar results and lead to similar conclusions about the goodness of fit of a given model, although chi-square is more robust for small samples.

EXAMPLE

For an example, Table 1 shows some data on voting behavior and class from 2 years in the British election of the late 1980s. The null hypothesis would suggest that the cell frequencies are randomly distributed. However, the chi-square for this is 84.7, and the G^2 (or likelihood ratio chi-square) is 88.4 for 4 *df*. These are clearly high values on a chi-square distribution based around

Table 1 Voting Data From Two British Elections in the Late 1980s, by Class

Year	Class	Vote		Total
		Labor	Conservative	
Year 1	Working class	50	20	70
	Middle class	10	40	50
Year 2	Working class	70	30	100
	Middle class	20	80	100
Total		150	170	320

4 *df* and, therefore, make it highly improbable that this model reflects reality. Constraining the model so that expected values of vote and class are expected to vary with each other, but so that the variation is constant across the 2 years, produces a model with Pearson and likelihood ratio chi-squares of 2.1 for 3 *df*. Clearly, such a value, when compared with the chi-square distribution having 3 *df*, is very probable ($p = .54$) and therefore represents a good fit with relative parsimony. The change in the chi-square and G^2 values (roughly 86 or 82 for 1 *df*) clearly represents a highly significant change.

HISTORY

The probabilities for a number of chi-square distributions were tabulated by Palin Elderton in 1901, and it was these tables that were used by Karl Pearson and his researchers in demonstrations of his chi-square test in the first decades of the 20th century. However, it is now widely acknowledged that Pearson's calculation of the numbers of degrees of freedom was misspecified, although Fisher's attempt to correct him was rejected and was only later acknowledged as being the correct solution. Pearson was therefore assessing his probability values from distributions that had too many degrees of freedom. Maximum likelihood ratios have been traced in origin to the 17th century, although it was the latter half of the 20th century that saw the real development and use of logit and log-linear models, as well as their accompanying GOODNESS-OF-FIT statistics, with the work of researchers such as Leo Goodman and Stephen Fienberg.

—Lucinda Platt

See also CHI-SQUARE TEST, DEGREES OF FREEDOM, LOG-LINEAR MODEL, LOGIT MODEL

REFERENCES

- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley.
- Fienberg, S. E. (1977). *The analysis of cross-classified categorical data*. Cambridge: MIT Press.
- Yule, G. U., & Kendall, M. G. (1950). *An introduction to the theory of statistics* (14th ed.). London: Griffin.

CHI-SQUARE TEST

The chi-square test is the most commonly used significance test for categorical variables in the social sciences. It was developed by Karl Pearson around 1900 as a means of assessing the relationship between two categorical variables as tabulated against each other in a CONTINGENCY TABLE. The test compares the actual values in the cells of the table with those that would be expected under conditions of independence (i.e., if there was no relationship between the variables being considered). Expected values are calculated for each cell by cross-multiplying the row and column proportions for that cell and taking as a share of the

total number of cases considered. For example, Table 1 shows the expected counts in a hypothetical example of voters, which have been obtained on the basis of the row and column totals (known as the marginal distributions) and the total number of cases. It also shows the observed counts, which can be seen to differ from those determined by independence. In this situation of independence, voting behavior does not vary with sex.

Of course, even where there is independence in the population, the observed values in a sample are very unlikely exactly to mimic the expected values. What the chi-square test ascertains is whether any differences between observed and expected values in the cells of the table according to the sample imply real differences in the population (i.e., that the counts are not independent). It does this by comparing the actual cell counts with the cell counts expected if the proportions were consistent across each category of the explanatory variable. The value of the chi-square statistic is calculated as

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e},$$

Table 1 Expected Voting Behavior by Sex Under Conditions of Independence

	Party A	Party B	Other Parties	Total
Male	$0.46 \cdot 0.5 \cdot 70,100$	$0.46 \cdot 0.3 \cdot 70,100$	$0.46 \cdot 0.2 \cdot 70,100$	
Expected	16,123	9,674	6,449	32,246 (46%)
Observed	13,669	13,811	4,766	
Female	$0.54 \cdot 0.5 \cdot 70,100$	$0.54 \cdot 0.3 \cdot 70,100$	$0.54 \cdot 0.2 \cdot 70,100$	
Expected	18,927	11,356	7,571	37,854 (54%)
Observed	21,381	7,219	9,254	
Total	35,050 (50%)	21,030 (30%)	14,020 (20%)	70,100 (100%)

Table 2 Difficulty Making Ends Meet by Social Class, Britain 1987 (expected counts in parentheses)

	Easy Making Ends Meet	Neither Easy Nor Difficult	Difficult Making Ends Meet	Total
Professional/managerial	124 (84)	76 (86)	57 (87)	257 26%
Intermediate	73 (75)	84 (77)	73 (73)	230 23.3%
Working class	125 (163)	171 (167)	204 (169)	500 50.7%
Total	322 32.6%	331 33.5%	334 33.8%	987 100%

where f_o = the observed cell count and f_e = the expected cell count. The test then takes into account the size of the sample and the DEGREES OF FREEDOM in the table to determine whether the differences in the sample are likely to be due to chance or the probability that they are reflected in the population (i.e., that they are significant). The degrees of freedom are calculated as (number of rows – 1) $\times$ (number of columns – 1). The size of the chi-square test statistic can then be related to the CHI-SQUARE DISTRIBUTION for those degrees of freedom to ascertain the probability that the divergence of the observed from the expected values could occur within that distribution or whether it likely falls outside it.

As an example, Table 2 shows the difficulty of making ends meet tabulated by class for a sample of 987 people living in Britain in the late 1980s. The chi-square statistic resulting from comparing the observed counts in each cell with the expected counts (given in brackets) is 47.9 for 4 degrees of freedom. This figure is significant at the .000 level and thus shows that the probability that financial difficulty and social class are independent is less than 1 in a thousand.

Chi-square does not provide information about the strength of an association or its direction, only about the probability of dependence between the variables. The value of chi-square increases with sample size, and thus a low p value (high significance) in a large sample may come with a fairly weak relationship between the variables. The chi-square test also does not indicate which values of the response variable vary with the explanatory variable. It may be that only one or two cell counts deviate greatly from their expected counts. This can be ascertained by examining the table in more detail. It is not appropriate to use the chi-square test when an expected cell count for any cell in the table is less than 5. This may occur with small samples or with a large number of cells in the table. In these situations, an alternative test, such as Fisher's exact test, may be more useful.

—Lucinda Platt

REFERENCES

- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley.
- Everitt, B. S. (1977). *The analysis of contingency tables*. London: John Wiley.
- Fienberg, S. E. (1977). *The analysis of cross-classified categorical data*. Cambridge: MIT Press.

CHOW TEST

In his article, "Tests of Equality Between Sets of Regression Coefficients in Linear Regression Models," econometrician Gregory Chow (1961) showed how the F test, or the VARIANCE RATIO TEST, which is widely used in ANALYSIS OF VARIANCE to test the homogeneity or equality of a set of means, can be used to find out the structural or parameter stability of REGRESSION models.

To explain the test, suppose we have data for the United States on savings (Y) and personal income (X), say, for the years 1970 to 1995, with Y and X being measured in billions of dollars.

Suppose we consider the following regression model, or the savings function:

$$Y_t = \beta_1 + \beta_2 X_t + u_{it}, \quad (1)$$

where Y is savings, X is income, u_{it} is the stochastic error term, and t is the time index.

One can use ORDINARY LEAST SQUARES (OLS) to estimate the parameters of the preceding savings regression. The question that the Chow test poses is this: Do the regression coefficients, β_1 and β_2 , remain stable over the entire sample period? There was a severe economic recession in the United States in 1981–1982. Has the recession changed people's savings behavior? To answer the question, suppose one divides the sample data into two periods, 1970–1981 and 1982–1995, and estimates two separate regressions as follows:

$$\text{Period 1970–1981: } Y_t = \alpha_1 + \alpha_2 X_t + u_{2t}, \quad (2)$$

$$\text{Period 1981–1995: } Y_t = \lambda_1 + \lambda_2 X_t + u_{3t}, \quad (3)$$

where the us are the time-specific error terms.

If the relationship between savings and income is stable over time, one would expect that $\beta_1 = \alpha_1 = \lambda_1$ (i.e., the intercepts are statistically the same) and $\beta_2 = \alpha_2 = \lambda_2$ (i.e., the slope coefficients are the same). But how does one find that out? Chow suggests the following steps: Estimate the three regressions given above and obtain their respective RESIDUAL SUMS OF SQUARES (RSS), $S_1(df = 24)$, $S_2(df = 10)$, and $S_3(df = 12)$, respectively. Note that in all, there are 26 observations: 12 in 1970–1981 and 14 in 1982–1995.

If it is assumed that the error terms u_2 and u_3 are normally distributed, have the same variance (i.e., $\text{var}(u_2) = \text{var}(u_3) = \sigma^2$), and are independently distributed, the Chow test proceeds as follows: The assumption that the error terms are independently distributed implies that the two samples (i.e., sample periods) are independent. Hence, add S_2 and S_3 and obtain what may be called the *unrestricted residual sum of squares* (RSS_{ur}), which will have $(n_1 + n_2 - 2k)df$, where n_1 and n_2 are the number of observations in the first and second periods, and k is the number of parameters estimated in the model (two in the present instance).

Because the RSS , S_1 , was obtained under the assumption that there is parameter stability (i.e., the regression coefficients do not change over the period), one can call S_1 the *restricted residual sum of squares* (RSS_r). This RSS will have $(n - k)df$, where $n = n_1 + n_2$. Now if there is parameter stability, RSS_{ur} and RSS_r should not be statistically different. But if there is no parameter stability, the two RSS s will differ. This can be seen, given the assumptions underlying the Chow test, by showing that

$$F = \frac{(\text{RSS}_r - \text{RSS}_{\text{ur}})/k}{(\text{RSS}_{\text{ur}})/(n_1 + n_2 - 2k)} \sim F_{[k, (n_1 + n_2 - 2k)]}.$$

That is, the F value thus computed follows the F distribution with the stated degrees of freedom. Therefore, if the computed F in an application exceeds, say, the 1%, 5%, or 10% level of significance, one may reject the hypothesis that there is parameter stability. In the present case, that would mean that the savings function has changed over time. On other hand, if the computed F value is not significant, one can assume that there is parameter stability. In this case, there may be no harm in running just one regression pooling all the observations.

Some numerical results of the Chow test can be shown from the data on savings and income that were obtained for the United States for 1970–1995, with savings and income being measured in billions of dollars. The various sums of residuals squares discussed previously were as follows:

$$S_1 = 23,248.30 \text{ (} df = 24 \text{)};$$

$$S_2 = 1785.032 \text{ (} df = 10 \text{)};$$

$$S_3 = 10,005.22 \text{ (} df = 12 \text{)}.$$

Therefore, in the present instance, $\text{RSS}_r = S_1 = 23,248.30$ and $\text{RSS}_{\text{ur}} = S_2 + S_3 = 11,790.252$. Inserting these values in the previous F ratio gives the following:

$$F = \frac{(23,248.30 - 11,790.252)/2}{(11,790.252)/22} = 10.69.$$

Under the assumed conditions, this F value follows the F distribution with 2 and 22 df in the numerator and denominator, respectively. The 1% critical F value (i.e., the probability of committing a TYPE I ERROR) is 5.72. Because the computed F value exceeds this critical value, we can say that the savings-income relation in the United States over the sample period has undergone a structural change: The savings function in the two subperiods is not the same.

In using the Chow test, it must be remembered that the assumptions underlying the test must be fulfilled and that the researcher must know when a change in the relationship might have occurred. In the preceding example, if one is not sure that the savings-income relationship changed in 1982, the Chow test cannot be applied. Of course, one can choose different break points and obtain the relevant F values, thereby choosing that break point that gives the largest F value. This modification of the Chow test is known as the *Quandt likelihood ratio statistic*.

—Damodar N. Gujarati

REFERENCES

- Chow, G. (1961). Tests of equality between sets of regression coefficients in linear regression models. *Econometrica*, 28(3), 591–605.
- Gujarati, D. (2002). *Basic econometrics* (4th ed.). New York: McGraw-Hill.
- Wooldridge, J. (2002). *Introductory econometrics* (2nd ed.). Cincinnati, OH: South-Western College Publishing.

CLASSICAL INFERENCE

To make an inference is to draw a conclusion from a set of premises through some acceptable form of reasoning. Inferential reasoning is the foundation on which the scientific method is based. Classical inferences are embedded in a deterministic framework and follow either a deductive or inductive approach. Other

forms of inference exist, but they are usually embedded in a probabilistic framework.

Within the classical framework, deductive reasoning draws a conclusion implicit in the stated premises. Inductive reasoning, however, argues from a series of specifics to a general statement. When well formulated, deductive arguments can be shown to be logically valid or “true.” Inductive reasoning, on the other hand, does not entail logical closure; thus, inductive conclusions cannot be shown, incontrovertibly, to be valid. This is why some logicians reject inductive reasoning as a legitimate form of logic.

Deductive reasoning is best exemplified by the classical syllogism that consists of a major premise, a minor premise, and a conclusion. Symbolically, we may write the following:

$$\begin{array}{l} \text{If } p, \text{ then } q \\ p \\ \hline \text{therefore, } q. \end{array}$$

A discursive example is as follows: If one is human, then one is mortal; Diana is human; therefore, Diana is mortal. This form of syllogism is known as the *modus ponens*. Another logically acceptable form of syllogism is the *modus tolens*, and it takes the following form: If p , then q ; $\sim q$; therefore, $\sim p$. Thus, we may state the following: If one is human, then one is mortal; Diana is not mortal; therefore, Diana is not human. In this instance, Diana might be the classical Greek goddess.

There are two other ways of writing the classical syllogism, but both entail errors in reasoning. The fallacy of affirming the consequent, for example, takes the following form: If p , then q ; q ; therefore, p . To continue with our substantive example: If one is human, then one is mortal; Diana is mortal; therefore, Diana is human. The error here is that being mortal is not a sufficient condition for being human. In fact, Diana might be one’s pet cat.

The second false conclusion, or the fallacy of denying the antecedent, takes the following form: if p , then q ; $\sim p$; therefore, $\sim q$. Again, discursively: If one is human, then one is mortal; Diana is not human; therefore, Diana is not mortal. Once more, Diana might be one’s cat, but although cats are not human, they do share the characteristic of mortality with humans. These latter two invalid forms are mistakes commonly made in scientific reasoning.

It is the logical form of the *modus tolens* that forms the backbone of the scientific hypothesis. Let us assume that hypothesis H implies some outcome O . Observing O , no matter how often, does not support the hypothesis H . In fact, hypotheses cannot be confirmed directly because using O to support H is consistent with the fallacy of affirming the consequent. Not observing O , however, provides a negation of H or a rejection of the hypothesis. Thus, as the philosopher Karl Popper noted, the hallmark of the scientific method is its ability to disprove, not prove, hypotheses.

Although deduction forms a closed system of logic, induction is an open system. Induction generally argues that because all observed instances of P have characteristic C , then *all* instances of P have characteristic C . As Hume so clearly noted, however, the statement that all observed A s are B s is not inconsistent with the statement that some unobserved A s are not B s. Proponents of induction argue that although things cannot be proven inductively, induction takes on more credence as the population becomes increasingly enumerated. Thus, the closer one comes to observing all instances of P where those instances have characteristic C , then the more credence we place in the statement that all instances of P have characteristic C . This rationale is appealing when P is a closed population; it is difficult to abide when P is open or infinite.

The difficulties surrounding logical inference do not necessarily imply that it has no use in science. Often, hypotheses are generated through processes of inference—something we call the context of discovery. It is also possible to draw credible inferences by generalizing from samples to populations outside a deterministic framework. This leads to the art known as INFERENTIAL STATISTICS, in which, given certain constraints, the likelihood of a valid inference can be estimated.

—Paul S. Maxim

REFERENCES

- Copi, I. M., & Cohen, C. (1990). *Introduction to logic* (8th ed.). New York: Macmillan.
- Gensler, H. J. (2002). *Introduction to logic*. London: Routledge Kegan Paul.
- Hempel, C. G. (1966). *Philosophy of natural science*. Englewood Cliffs, NJ: Prentice Hall.
- Poundstone, W. (1988). *Labyrinths of reason: Paradoxes, puzzles and the frailty of knowledge*. New York: Anchor/Doubleday.

CLASSIFICATION TREES. See CART

CLINICAL RESEARCH

Clinical research is inquiry whose purpose is the creation of new knowledge and action in the context of health care. As with all forms of scientific inquiry, clinical research concerns itself with identification, description, the generation of EXPLANATIONS, the testing of explanations, and/or intervention/CONTROL. These concerns, in turn, are linked to typical methods. For the first three—exploratory investigations—qualitative methods of OBSERVATIONS, INTERVIEWS, and DOCUMENT analysis are commonly used, as are variations of surveys. Explanation testing, although part of qualitative traditions, is more commonly practiced in epidemiological research, with COHORT, CROSS-SECTIONAL, and case control designs. For example, an EXPERIMENTAL design called a RANDOMIZED CONTROL TRIAL (RCT) tests the efficacy or effectiveness of treatment (Miller & Crabtree, 1999). The health disciplinary literature, particularly in medicine, is still dominated by experimental and epidemiological designs. This distribution, in turn, reflects funding priorities for basic and applied biomedical research in developed countries. Similar models also dominated social sciences relevant to clinical practice, such as psychology and sociology, in the early years of these disciplines. However, there are now established qualitative and critical streams in these and other social sciences, and new inquiry models are applied to the questions of the clinic. In an era of “evidence-based medicine,” qualitative investigators remind us that the care of patients must include many kinds of evidence—those that derive from randomized control trials and those that derive from being acquainted with the particulars of patients’ lives, perspectives, and social context (Morse, Swanson, & Kuzel, 2001).

One of these newer models of clinical inquiry is PHENOMENOLOGY, which focuses on understanding and communicating the essence of experiences and intentions of individuals within their “life-world.” In contrast to disease-oriented descriptions, phenomenological reports portray the illness experience. For example, Pelusi (1997) interviewed breast cancer survivors and found common themes of not only loss and an uncertain future but also growth and enlightenment. Another is GROUNDED THEORY, well suited to the

evaluation of social processes over time, with a goal of generating theory that is grounded in careful observation, interview, and document analysis. An example is Lorencz’s (1991) study of patients with schizophrenia who transition from inpatient to community settings—a process of “becoming ordinary.” Anthropology’s basic tradition of ETHNOGRAPHY has also seen wide application to the work of the clinic. This includes not only the cultural determinants of health and illness (Kleinman, Eisenberg, & Good, 1978) but also the culture of health care organizations and its impact on the delivery of clinical services. A high-profile application of an ethnographic sensibility is to the widespread problem of “medical error,” resulting in a growing appreciation that a culture of “blame and shame” is a powerful deterrent to identifying and understanding the root causes of medical errors (Kohn, Corrigan, & Donaldson, 2000). A final example is PARTICIPATORY ACTION RESEARCH, with an interest in understanding and promoting social change through the “democratization” of inquiry. In this model, those who would typically be “subjects” of medical research act instead as coinvestigators. Macaulay and colleagues (1997), in a Mohawk community near Montreal, focused on diabetes prevention through healthy eating and more exercise among children. They employed both behavioral change models and knowledge of native learning styles to design interventions.

—Anton J. Kuzel

REFERENCES

- Kleinman, A. L., Eisenberg, L., & Good, B. J. (1978). Culture, illness and care: Clinical lessons from anthropologic and cross-cultural research. *Annals of Internal Medicine*, 88, 251–258.
- Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (Eds.). (2000). *To err is human: Building a safer health system*. Washington, DC: National Academy Press.
- Lorencz, B. (1991). Becoming ordinary: Leaving the psychiatric hospital. In J. M. Morse & J. Johnson (Eds.), *The illness experience: Dimensions of suffering* (pp. 140–200). Newbury Park, CA: Sage.
- Macaulay, A. C., Paradis, G., Potvin, L., Cross, E. J., Saad-Haddad, C., McComber, A., et al. (1997). The Kahnawake Schools Diabetes Prevention Project: Intervention, evaluation, and baseline results of a diabetes primary prevention program with a native community in Canada. *Preventive Medicine*, 26(6), 79–90.
- Miller, W. L., & Crabtree, B. F. (1999). Clinical research: A multimethod typology and qualitative roadmap. In B. F. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (2nd ed., pp. 3–32). Thousand Oaks, CA: Sage.

Morse, J. M., Swanson, J. M., & Kuzel, A. J. (Eds.). (2001). *The nature of qualitative evidence*. Thousand Oaks, CA: Sage.
 Pelusi, J. (1997). The lived experience of surviving breast cancer. *Oncology Nursing Forum*, 24(8), 1343–1353.

CLOSED QUESTION. See
 CLOSED-ENDED QUESTIONS

CLOSED-ENDED QUESTIONS

Survey questions come in two varieties: OPEN-ENDED QUESTIONS, in which the respondents provide their own answers, and closed-ended questions, in which specific response categories are provided in the question itself. Although there has been considerable research on the relative merits of the two types of questions, the substantial preponderance of questions that appear in any survey are closed-ended questions.

All survey research involves asking questions for which the responses are categorized to facilitate analysis. In closed-ended questions, the researcher makes prior judgments about what the appropriate categories might be and offers them immediately to the respondent in the question wording. There are a number of potential problems with this format, not the least of which is that it typically describes the “world” in dichotomous terms that sometimes reflect an oversimplification of possibilities. This kind of constraint does not exist with an open-ended question, and for this reason Schuman and Presser (1996) conclude that they often provide more valid data. However, the cost saving of not having to CODE verbatim responses and the quicker access to analysis make close-ended questions the preferred form.

When designing or using close-ended questions, a number of standard suggestions reflect considerable research on the matter. Using forced choices is preferable to asking a respondent to agree or disagree with a single statement. A middle alternative should generally be offered, except when measuring intensity, and an explicit “no opinion” option should be offered as well. Researchers should use multiple questions to assess the same topic, remaining sensitive to the effects of question order. And when in doubt about the possible effects of question wording or response options, research should include split-sample versions in their questionnaires so that comparative analyses can be run.

—Michael W. Traugott

REFERENCES

- Converse, J. M., & Presser, S. (1986). *Survey questions*. Thousand Oaks, CA: Sage.
 Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: Sage.
 Schuman, H., & Presser, S. (1996). *Questions & answers in attitude surveys*. Thousand Oaks, CA: Sage.

CLUSTER ANALYSIS

Social science DATA sets usually take the form of observations on UNITS OF ANALYSIS for a set of VARIABLES. The goal of cluster analysis is to produce a simple classification of units into subgroups based on information contained in some variables. The vagueness of this statement is not accidental. Although there may be no formal definition of cluster analysis, a slightly more precise statement is possible. The *clustering problem* requires solutions to the task of establishing clusterings of the n units into r clusters (where r is much smaller than n) so that units in a cluster are similar, whereas units in distinct clusters are dissimilar. Put differently, these clusterings have homogeneous clusters that are well separated. *Cluster analysis* is a label for the diverse set of tools for solving the clustering problem (see Everitt, Landau, & Leese, 2001). Most often, these tools are used for INDUCTIVE explorations of data. The hope is that the clusterings provide insight into the structure of the data, the nature of the units, and the processes generating the variables. For example, cities can be clustered in terms of their social, economic, and demographic characteristics. People can be clustered in terms of their psychological profiles or other attributes they possess.

DEVELOPMENT OF CLUSTER ANALYSIS

Prior to 1960, many clustering problems were solved separately in different disciplines. Progress was fragmented. The early 1960s saw attempts to provide general treatments of cluster analysis, given these many developments. Sokal and Sneath (1963) provided an extensive discussion and helped set the framework for the development of cluster analysis as a data-analytic field. Specifying clustering problems is not difficult. Nor are the mathematical foundations for expressing and creating most solutions to the clustering problem. The difficulty of cluster analysis comes

from the *computational complexities* in *establishing* solutions to the clustering problem. As a result, the field has been driven primarily by the evolution of computing technology. Generally, this has been beneficial, with substantive interpretations being enriched by useful clusterings. In addition, many technical developments have stemmed from exploring substantive applications in new domains. There are now many national societies of cluster analysts that are linked through the International Federation of Classification Societies.

SOLVING CLUSTERING PROBLEMS

In general, the clustering problem can be stated as establishing one (or more) clustering(s) with r clusters that have the minimized value of a well-defined criterion function over all feasible clusterings. The criterion function provides a measure of fit for all clusterings. In practice, however, the criterion function often is left implicit or ignored. In most applications, the clustering is a partition, but “fuzzy clustering” with overlapping clusters is possible. Once the units of analysis have been selected, there are five broad steps in conducting cluster analyses:

1. measuring the relevant variables (both QUANTITATIVE VARIABLES and CATEGORICAL variables can be included, and some form of standardization may be necessary),
2. creating a (dis)similarity MATRIX for an appropriate measure of (dis)similarity,
3. creating one or more clusterings via a clustering algorithm,
4. providing some assessment of the obtained clustering(s), and
5. interpreting the clustering(s) in substantive terms.

Although all steps are fraught with hazard, Steps 2 and 3 are the most hazardous, and Step 4 is ignored often. In Step 2, dissimilarity measures (e.g., Euclidean, Manhattan, and Minkowsky distances) or similarity measures (e.g., CORRELATION and matching COEFFICIENTS) can be used. The choice of a measure is critical: Different measures can lead to different clusterings. In Step 3, there are many ALGORITHMS for establishing clusterings. Each pair of choices (of measures and algorithms), in principle, can lead to different clusterings of the units.

Hierarchical clustering can take an agglomerative or a divisive form. An agglomerative clustering starts with each unit in its own cluster and systematically merges units and clusters until all units form a single cluster. Divisive hierarchical clustering proceeds in the reverse direction. Within these categories, there are multiple options. Three of the most popular are single-link, average-link, and complete-link clustering. In single-link clustering, the algorithm computes the (dis)similarity between groups as the (dis)similarity between the closest two units in the two groups. For complete-linkage clustering, the farthest pair of units in the two groups is used, and in average-linkage clustering, the algorithm uses the average (dis)similarity of units between the two groups. Ward’s method is popular also for computing ways of combining clusters. The choice between these methods is critical, as shown in Figure 1.

The first panel of Figure 1 shows a BIVARIATE scattergram whose 25 units can be classified. Squared Euclidean distance was used for all three clusterings and dendrograms shown below. For this example, if (x_i, y_i) and (x_j, y_j) are the coordinates of two units, the squared Euclidian distance between them is $(x_i - x_j)^2 + (y_i - y_j)^2$. The top right panel shows the single-linkage clustering dendrogram. Those in the bottom row of Figure 1 are for the average- and complete-linkage methods. The scale on the left of the dendrogram shows the measure of dissimilarity, and the horizontal lines show when units or clusters are merged. The vertical lines keep track of the clusters. The average- and complete-linkage clusterings are the closest to each other, and both suggest clusterings with three clusters. However, the clusters differ in their details, and the single-link clustering differs markedly from the other two. Which clustering is “better” can be judged by examining the clusterings and the scattergram—with some idea of how and why units can be grouped. For real analyses, substance guides this judgment.

Because clustering tools are exploratory, there are few tools for STATISTICAL INFERENCE concerning the VALIDITY of a clustering (Step 4), and their utility is limited because of the prior exploration. Choosing the number of clusters from a dendrogram often is viewed as a matter of judgment, and ad hoc justifications for any clustering are easy to reach given a clustering. This, too, is illustrated in Figure 1.

When the number of clusters is known or assumed, some nonhierarchical clustering methods are available.

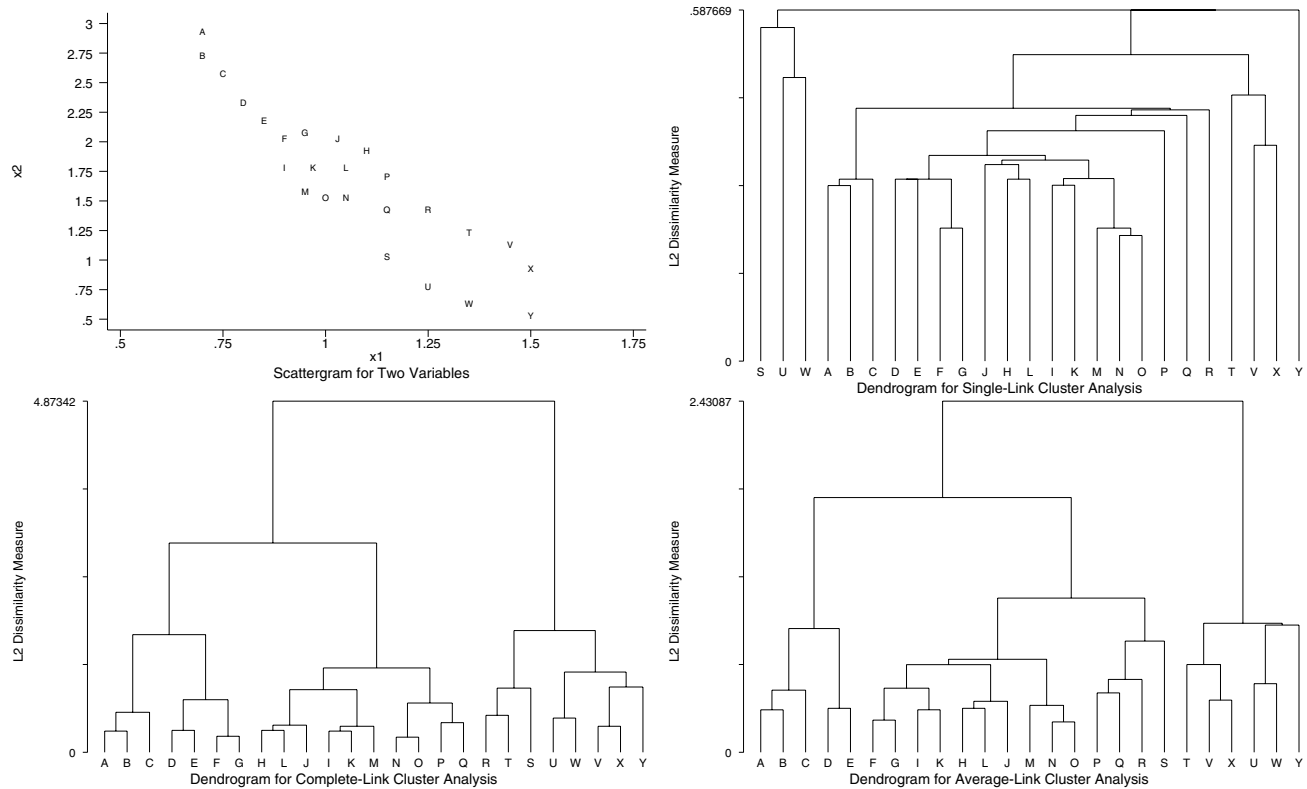


Figure 1 Three Clusterings of a Common Data Set

These include *k*-means and *k*-medians clustering. Two additional methods are the leader and relocation algorithms. Both are local optimization methods and have to be repeated many times to avoid reaching a local minimum. The relocation algorithm is at the heart of the direct clustering approach to block modeling social networks developed by Doreian, Batagelj, and Ferligoj (1994). Both network actors *and* the network ties are clustered, whereby a criterion function is defined *explicitly* in terms of substantive concerns and the NETWORK ties. As a result, the resulting solutions to the clustering problem are not ad hoc.

—Patrick Doreian

REFERENCES

- Doreian, P., Batagelj, V., & Ferligoj, A. (1994). Partitioning networks based on generalized concepts of equivalence. *Journal of Mathematical Sociology*, 19, 1–27.
- Everitt, B., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). Oxford, UK: Oxford University Press.

Sokal, R., & Sneath, P. (1963). *Principles of taxonomy*. San Francisco: Freeman.

CLUSTER SAMPLING

Cluster sampling involves sorting the units in the study population into groups and selecting a number of groups. All the units in those groups are then studied. It is a special case of MULTISTAGE SAMPLING.

—Peter Lynn

COCHRAN'S Q TEST

W. G. Cochran (1950) developed the *Q* statistic for matched or WITHIN-SUBJECT DESIGNS in which each subject (*r*) provides a dichotomous response for each experimental condition (*c*). Cochran's *Q* tests whether the probability of a target response is equal across

Table 1 Sample Data for the Cochran's Q : Tests of Varying Levels of Complexity

	Test 1	Test 2	Test 3	Test 4	
Subject 1	1	0	0	1	$u_1 = 2$
Subject 2	1	1	1	0	$u_2 = 3$
Subject 3	1	0	0	0	$u_3 = 1$
Subject 4	1	1	1	1	$u_4 = 4$
Subject 5	1	0	1	0	$u_5 = 2$
Subject 6	1	1	0	0	$u_6 = 2$
	$T_1 = 6$	$T_2 = 3$	$T_3 = 3$	$T_4 = 2$	$\bar{T} = \frac{\sum_{j=1}^c T_j}{c} = 3.5$
$Q = \frac{c(c-1) \sum_{j=1}^c (T_j - \bar{T})^2}{c(\sum_{i=1}^r u_i) - (\sum_{i=1}^r u_i^2)} = \frac{4(3)[(6-3.5)^2 + (3-3.5)^2 + (3-3.5)^2 + (2-3.5)^2]}{4(2+3+1+4+2+2) - (2^2 + 3^2 + 1^2 + 4^2 + 2^2 + 2^2)} = 6$					
95th percentile for $\chi^2_{(4-1)} = 7.815$					

all conditions. For example, if 6 subjects ($r = 6$) take four tests ($c = 4$) that are either “failed” or “passed,” Cochran's Q can assess whether the probability of passing differs across the four tests. Of course, the dependent variable can be any dichotomy, such as pass-fail, presence-absence, hit-miss, or increase-decrease.

Cochran's Q can easily be computed with data arranged in a matrix of r rows and c columns. Each row in the matrix represents a case containing c elements that are coded as “1” or “0” to represent the presence or absence of a target response. The number of target responses are summated across rows (u_i) and columns (T_j) and are used to compute Cochran's Q in the following equation:

$$Q = \frac{c(c-1) \sum_{j=1}^c (T_j - \bar{T})^2}{c(\sum_{i=1}^r u_i) - (\sum_{i=1}^r u_i^2)},$$

where T_j is the total of the j th column, u_i is the total of the i th row, and $\bar{T} = \frac{\sum_{j=1}^c T_j}{c}$.

Cochran's Q is distributed as a χ^2 with $c - 1$ DEGREES OF FREEDOM. As shown in Table 1, data from 6 subjects who each took four tests are analyzed with the Cochran's Q . The observed Q statistic is then compared to a χ^2 using the desired TYPE I ERROR rate (.05 in this case) and degrees of freedom. In the example, the Q statistic does not exceed the critical χ^2 . Thus, the NULL HYPOTHESIS is retained, suggesting that the probability of passing the test does not differ across conditions.

Cochran's Q requires the assumption that cases are randomly and independently sampled from

the population. Homogeneity of covariance is also required; the covariances of all possible pairs of columns are assumed to be equal. If the assumption of homogeneity of covariance is violated, Myers and Well (1995) recommended using the ε -adjustment developed by Greenhouse and Geisser (1959). The corrected Q is the following: $Q^* = \varepsilon Q$.

Degrees of freedom for the adjusted test are $\varepsilon(c-1)$. As before, Q^* is evaluated against the desired χ^2 with the adjusted degrees of freedom. The computation of ε is laborious and is omitted here. However, it is given by many statistical analysis programs when conducting repeated-measures ANALYSIS OF VARIANCE.

In using the Cochran's Q , sample sizes of 16 or more are advisable (the sample size in the table is only for illustration). Cases that exhibit no variance across conditions (i.e., all zeros or ones) do not contribute to Q . The repeated-measures analysis of variance can be considered as an alternative to Q , although large samples are needed to ensure normality of the sampling distribution of the dependent variable, and the ε -adjustment may be needed to correct for heterogeneity of covariance.

—James W. Griffith

REFERENCES

- Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256–266.
- Greenhouse, G. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 55, 431–433.
- Myers, J. L., & Well, A. D. (1995). *Research design and statistical analysis*. Hillsdale, NJ: Lawrence Erlbaum.

CODE. See CODING

CODEBOOK. See CODING FRAME

CODING

Coding is the process by which verbal data are converted into VARIABLES and categories of variables using numbers, so that the data can be entered into computers for analysis. DATA for social science research are collected using SELF-ADMINISTERED QUESTIONNAIRES and questionnaires administered through telephone or face-to-face interviews, through observation, and from records, documents, movies, pictures, or tapes. In its original “raw” form, these data comprise verbal or written language or visual images. Although many data entry programs can handle at least limited amounts of linguistic data, for purposes of most analyses, these data must eventually be converted to variables and numbers and entered into machine-readable data sets. Each variable in a data set (e.g., sex or gender) must consist of at least two categories, each with a unique code, to be a variable. It must have the possibility of varying; if it does not vary, then it is a constant. Thus, a study could include both men and women, where gender or sex is a variable, or it could include only men or only women, where gender is a CONSTANT. Numeric codes for gender can be assigned in both situations.

A single unique code can consist of a single digit or multiple digits, with the functional upper number of digits being 8 or 10. A code consisting of multiple digits may have multiple variables embedded within it. For example, respondents were asked after earthquakes whether various utilities were unavailable, including water, electricity, gas, and telephones. Four two-category variables can be created for each utility, using 1 for “yes, it was off” and 2 for “no, it was not off.” But the four single-digit variables can also be combined to create a four-digit variable that creates all the possible combinations of the four variables. An example would be the code of 1221, which would indicate that water and telephones went off but electricity and gas stayed on.

The numbers assigned can be meaningful in and of themselves, or they can function simply as a shorthand representation of the verbal data. If, for example, data

on age are collected, the age recorded for each person, institution, state or country, or object is meaningful, and statements can be made about the relative ages of people, states, or objects. If, in contrast, data on religious affiliations of people, households, or institutions are collected, the number assigned to each religion is arbitrary—for example, 1 for Catholics, 2 for Buddhists, and 3 for Hindus.

LEVELS OF MEASUREMENT

In developing codes for verbal data, the concept of LEVELS OF MEASUREMENT is often used (Stevens, 1968). There are four levels of measurement: NOMINAL, ORDINAL, INTERVAL, and RATIO. As we move from nominal data to ratio data, the computations and statistics that can be used during analysis become more diverse and sophisticated. When creating codes for data, it is important to understand and consider these differences.

Codes for religion are considered nominal. Nominal codes are arbitrary, shorthand space holders. In creating codes for nominal data, the researcher is concerned with creating exhaustive and mutually exclusive codes. *Exhaustive* means that a unique code number has been created for each category of the variable that may occur in the data. Thus, in creating codes or code frames for a variable on religious affiliation, it may be necessary to create unique code numbers for individuals, institutions, or groups that have no religious affiliation, as well as for those that identify as agnostics or atheists. *Mutually exclusive* means that the information being coded about each person, institution, object, or grouping can be clearly assigned to only one category of the variable. So, for example, it is clear that Buddhists always get a code of 2 and never get a code of 1 or 3.

The categories of a nominal variable have no meaningful rank order, and the distance between the categories cannot be calculated. The only measure of CENTRAL TENDENCY that can be calculated appropriately for a nominal variable is the mode. Many of the most important variables included in social science research are naturally nominal. These include sex or gender, religion, language, ethnic identification, and marital status.

Ordinal variables are considered a higher level of measurement. The categories of an ordinal variable can be meaningfully rank ordered. So, for example, a person who drinks one pint of water drinks less water than a person who drinks two quarts who, in turn,

drinks less water than a person who drinks two gallons, and they can be assigned codes, respectively, of 1, 2, and 3. The person with a code of 3 clearly drinks more water than those with codes 1 or 2, but nothing can be said about the relative distance between the three people when those three codes are used in analysis. Both modes and MEDIANs can be calculated for ordinal variables, but it is technically incorrect to add or subtract scores or to calculate MEAN scores. One of the problems with assigning codes to variables that are naturally ordinal is developing codes that are both mutually exclusive and exhaustive. In the above example, the codes are mutually exclusive but are not exhaustive. Note there is no code for a person who drinks no water, three gallons of water, or one quart of water.

Age can be regarded as an instance of an interval variable, which requires a scale with equidistant (or constant) numerical categories. Consider the answer to the question, "How old were you on your last birthday?" With interval variables, the categories can be rank ordered, and one can talk about the equal distance between the categories. For instance, interval variables have an arbitrary zero, the categories can be rank ordered, and we can talk about the distance between categories. For example, the distance between 24 and 26 is the same as the distance between 33 and 35. In addition to the mode and median, statistics such as the mean, STANDARD DEVIATION, and CORRELATION are meaningful and can be interpreted.

Ratio variables differ from interval variables in that they have a meaningful zero and are completely CONTINUOUS, meaning that every point on the scale exists. Weight and height are often considered ratio variables. Ratio variables are also called continuous variables, whereas the other three types of variables are sometimes called DISCRETE variables. For purposes of most social science research, the distinction between interval and ratio is largely arbitrary because most of the statistical analyses that can be done on ratio variables can also be done on interval variables. In contrast, analysis techniques that are appropriate for interval and ratio variables technically should not be used on nominal or ordinal variables.

DETERMINING CODES FOR VARIABLES

Codes for data can be developed during the design of the study or after the data are collected. When structured questionnaires or abstraction forms are used to collect data on a finite number of variables and

most variables have a finite number of possible answer categories, codes should be developed as the data collection instrument is being finalized. When OBSERVATIONS, FOCUS GROUPS, SEMI-STRUCTURED INTERVIEWS, or DOCUMENTS are the source of data, codes generally will be developed inductively using CONTENT ANALYSIS or other procedures after data collection is completed. The development of codes for CLOSED-ENDED QUESTIONS in a questionnaire will be used as an example for the first kind of code development, whereas the development of codes for OPEN-ENDED QUESTIONS will be used as an example for the second kind.

A number of things should be considered when developing codes. As noted above, codes and the associated answer categories should be both *exhaustive* and *mutually exclusive*; that is, there should be a code for each answer that might be given, and RESPONDENTS, data collectors, and data entry personnel should have no difficulty determining which answer fits the situation being described and which number or code is associated with that answer. For example, after the Loma Prieta earthquake, the authors (Bourque & Russell, 1994; Bourque, Shoaf, & Nguyen, 1997) asked respondents, "Did you turn on or find a TV or radio to get more information about the earthquake?" and provided the responses of "Yes, Regular TV," "Yes, Battery TV," "Yes, Regular Radio," "Yes, Battery Radio," and "No," with preassigned codes of 1, 2, 3, 4, and 5. Interviewers asked respondents to select a single answer.

Although the list of answers and codes appeared both exhaustive and mutually exclusive, it was not. First, many respondents sought information from more than one form of electronic media and, second, many respondents sought information from their car radio. The question-and-answer categories were changed to add car radios to the list, and respondents were allowed to select more than one answer. If that had not been done, data would have been lost.

When multiple information is recorded or coded for what is technically one variable, each response category becomes a variable with two codes or values, mentioned or not mentioned. In the example, what started out as one variable with five possible answers became six interrelated variables, each with two possible answers or codes.

Residual other categories provide another way to ensure that the CODE FRAME is exhaustive. "Residual other" is used to deal with information that the researcher did not anticipate when the code frame was

created. In the same study, the author provided a list of ways in which houses might have been damaged by the earthquake, including collapsed walls, damaged roofs, and so forth. What was not anticipated were the elevated water towers that are used by residents of the hills outside Santa Cruz, California. Because a residual other was included, when respondents said a water tower had been damaged, interviewers circled the code of 19 for “Other” and wrote “water tower” in the space provided, as demonstrated here:

Other..... 19
SPECIFY: _____

When data are entered into the computer, both the code 19, which indicates that an unanticipated answer was given, and the verbal description of that answer, water tower, are typed into the data set.

Reviews of past research and PRETESTING of data collection procedures can help researchers determine if the answer options and associated codes are both mutually exclusive and exhaustive, but sometimes social change will cause well-tested code schemes to become inaccurate, incomplete, or even completely obsolete. For example, the standard categories used to code marital status are “never married,” “married,” “divorced,” “separated,” and “widowed,” but increasingly, both different- and same-sex couples are cohabiting. According to the 2000 U.S. census, 5.2% of households in the United States currently describe themselves as “unmarried partner households.” In a study of social networks, for example, failure to consider and account for this change may result in serious distortion of data.

Deciding on the categories and codes to use for income, education, and age is another place where problems occur. Generally, income is coded as an ordinal variable. Categories are developed such as the following: less than \$10,000, \$10,000–\$19,999, \$20,000–\$29,999, and so forth. The problem comes in deciding how broad or narrow each category should be and what the highest and lowest categories should be. If categories are too broad, the top category set too low, or the bottom category set too high, heaping can occur. *Heaping* is when a substantial proportion of the data being coded falls into a single category. For example, if undergraduate college students are being studied and age is set in ordinal categories of 18–21, 22–24, and 25–29 with associated codes of 1, 2, and 3, the overwhelming majority of students will be 18 to 21, and they will be assigned a code of 1. When this

happens, the variable becomes useless because it no longer has any VARIANCE.

Another problem with creating ordinal variables out of data (e.g., age, education, and income) that are naturally at least interval in underlying structure is that it may restrict data analysis. If age is coded into categories of 18–24, 25–29, 30–34, and 35–39, and so forth, the researcher can determine what the median age category is but can never calculate the mean age. Furthermore, writing out the age categories in a questionnaire or data collection form takes substantially more room than simply asking for age at last birthday or birth date. In this case, a *short open-ended* question in a data collection form may obtain just as much data—data that are just as valid and, at the same time, yield data that are more easily analyzed.

Similar problems occur when researchers use answer categories such as “none,” “some,” and “a lot.” Although *none* is pretty clear, what do *some* and *a lot* mean, and how should they be coded? If this is a question about how many people a respondent knows who were injured in a flood, why not ask instead, “How many people in all do you know who were injured in the flood?” The information collected will result in data that more closely resemble interval data, and it will not be subject to the variety of ways in which terms such as *some* and *a lot* will be interpreted.

When coding data, it is as important to record that information is missing as it is to record the data that are there. Standard MISSING DATA conditions include “not applicable,” “don’t know,” “refused,” and “missing.” Codes for “don’t know” and “refused” are more frequently used when data are collected directly from people, whereas codes for “missing” are extremely important when data are being abstracted out of records or other documents. “Not applicable” codes are used in both situations. “Not applicable” is appropriately used when some data are not relevant in certain situations. Often, the need for “not applicable” occurs because the data being coded are dependent on some prior information that results in the current data being irrelevant. For example, men are not asked about how they felt when they were pregnant, and it is irrelevant to look for hemoglobin scores in medical records if no blood was drawn.

Often, when data are being abstracted out of documents or records, data instructors are instructed to look for particular kinds of information and then record it using a code. If CONTENT ANALYSIS is being used to code the content of the front page of eight major

U.S. newspapers over a period of a month, and data collectors are supposed to record all articles that reference the president and code for the “tone” of the article (e.g., whether it is complimentary, critical, analytical, or whatever), the fact that no reference to the president occurs on a given front page is important information. Hence, a code for missing is included in the code frame or range of numeric codes being used to represent the content of the page. The code for missing provides at least indirect evidence that the data collector looked for information to code and did not find it.

The timing of and extent to which codes for “don’t know” should be used are one of the more controversial areas of data collection. The meaning and value of including a “don’t know” category and code differs with the following factors: the researcher’s objectives, how data are collected, whether factual or attitudinal information is being solicited, whether respondents are told that a “don’t know” category is available, and the extent to which data collectors are trained and monitored. When data are collected directly from people and the primary objective is to collect factual or behavioral information, it is better not to include a “don’t know” in the available answer categories during data collection because interviewers and respondents will be tempted to select “don’t know” when other answers are more appropriate. Complicating this is the fact that in interactions such as interviews, people often use phrases such as “don’t know” while they think about a question. In such cases, “don’t know” is an acknowledgment that the person heard the question, but it is not necessarily intended to be the answer to the question. When attitudes or opinions are being solicited, an identified “don’t know” or “no opinion” answer category may be appropriate, particularly when respondents are asked their opinion about people, policies, or practices that many know nothing about.

Some categories of answers occur repeatedly throughout data collection, such as yes, no, refused, don’t know, missing information, and inapplicable. Researchers can simplify data analysis by using *consistent numeric codes* throughout the data collection process for these common response options. Table 1 shows the traditional codes generally used.

Finally, researchers should consider developing code frames that *minimize the need for transformations* during data analysis. For example, if data about immunizations are being collected off medical records that represent well-baby visits, it makes much more sense to record the actual number of immunizations given

Table 1 Examples of Consistent Codes

Response	Example Code		
	Single Digit	Double Digit	3 + Digits
Yes	1		
No	2		
Refused to answer	7	97	997
Don’t know	8	98	998
Missing information	9	99	999
Not applicable	0	00	000

at each visit rather than creating categories for “no immunizations,” “1–2 immunizations,” “3–4 immunizations,” and “5 or more immunizations” with codes, respectively, of 1, 2, 3, and 4.

POSTCODING

Code frames for some kinds of data simply cannot be set up as the data collection procedures are being finalized. Open-ended questions that have no predetermined list of answer categories are included in interviews for a reason—namely, because the researcher is unable to anticipate the kinds and amounts of data that may be elicited or cannot create exhaustive, mutually exclusive codes that can easily be explained to and used by data collectors and data entry personnel. This is particularly true of exploratory research in which the objective is to find out whether more structured research is possible and would be of value. In such studies, interviews or other methods of data collection may be unstructured or only partially structured. Often, material is audio or visually recorded and transcribed prior to coding and analysis. When data collection is complete, the task is to determine what, if any, patterns exist in the data. This is done inductively through procedures such as content analysis. Similar procedures must be used to analyze documents, movies, and television programs. But the end objective is to come up with code frames or numeric representations of the data such that they can be analyzed and summarized.

Content analysis and related procedures are complex, and detailed information about how to do them is beyond the scope of this entry. Rather, some general guidelines can be provided here. First, the researcher must specify the objectives for which the code frame is to be used. If children have been observed at play, is

the objective to count the frequency of behaviors, the duration of behaviors, the intensity of the interactions, or some combination? By identifying how the data are to be used, the researcher identifies some variables that will be extracted and others that will be ignored. The process of identifying the objectives may involve preliminary gross content analysis so that overall themes can be inductively developed from the data.

Second, in developing the code frame, a balance must be realized between too much detail and not enough detail. If clothing is being described, is it important to code the number and type of buttons on the clothing, or is it sufficient to say it is a jacket? Third, and related, is maximizing the maintenance of information within the code frame as it is created. Of particular concern here is information that appears in only a few of the documents being analyzed but that is particularly pertinent to the research questions being examined. Because content analysis is a lengthy, labor-intensive process, researchers sometimes fail to code information that is important because they are tired or bored. As long as the raw data exist, the researcher can always reexamine the data, but this often means reading many transcripts or listening to many audiotapes. Thus, every effort should be made to incorporate rare but important data into the code frame as they are being developed. This is why actively searching for information and systematically coding the fact that certain information is missing is particularly important.

Finally, researchers need to create a sufficient range of codes, variables, or dimensions so that the coder does not force data into categories that are really inappropriate. Take, for example, answers to the following statement: "Describe the vehicle(s) involved in the motor vehicle crash." First, we could code for the number of vehicles that were involved in the crash. Then, we could code for the type of vehicle(s)—sedan, sport utility vehicle (SUV), or pickup truck—that was involved. If examining safety features is an objective of the study, it probably is important to code for the make of the vehicle, the year, and whether the vehicle had air bags. The point here is that a single stimulus—whether it is a question, a paragraph in a document, a period of observation, or a sequence in a movie or tape—may yield multiple pieces of information that become multiple variables during analysis.

—Linda B. Bourque

See also CODING QUALITATIVE DATA

REFERENCES

- Bourque, L. B., & Clark, V. A. (1992). *Processing data: The survey example* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-085). Newbury Park, CA: Sage.
- Bourque, L. B., & Russell, L. A. (with Krauss, G. L., Riopelle, D., Goltz, J. D., Greene, M., McAfee, S., & Nathe, S.) (1994, July). *Experiences during and responses to the Loma Prieta earthquake*. Oakland: Governor's Office of Emergency Services, State of California.
- Bourque, L. B., Shoaf, K. I., & Nguyen, L. H. (1997). Survey research. *International Journal of Mass Emergencies and Disasters*, 15, 71-101.
- Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science*, 161, 849-861.

CODING FRAME

A coding frame, code frame, or codebook shows how verbal or visual data have been converted into numeric data for purposes of analysis. It provides the link between the verbal data and the numeric data and explains what the numeric data mean.

At its simplest, a code frame refers to the range of numeric codes that are used in CODING information that represents a single VARIABLE (e.g., the state where people were born). In this example, at least 50 unique numbers would be included in the code frame, one for each state. Numbers would be assigned to states arbitrarily. Unique numbers would also be set aside for people who were not born in the United States and for people for whom no DATA were available. Once the data are entered into a data file, the code frame provides the only link between the original verbal data and the location and meaning of information in the data set. Without the code frame, the analyst will not know what the number 32 means or stands for and may not know that it represents a state.

At its most complex, a code frame or codebook shows where all data are located in a data file, what sets of numbers are associated with each piece of data or variable, and what each unique number stands for. Thus, in addition to state of birth, the data set might also include county of birth and date of birth. The code frame would describe the order in which these three variables were found in the data set, as well as what each unique code represented and the way in which date of birth was entered into the data set.

A code frame can be created prior to actual data collection, or it can be constructed afterwards through a process of CONTENT ANALYSIS and postcoding. When the number of variables and the range of probable data that will be collected for each variable are known in advance, it is advisable to precode the variables and thus create a code frame before actually collecting the data. If data are being collected through computer-assisted telephone interviews (CATI) or similar methods, code frames for most of the data must be prespecified or precoded to program the computer for data collection. A major advantage of CATI is that interviewers automatically code the data as respondents answer the questions. If the interview is dominated by questions that require interviewers to type in lengthy text answers, the value of using CATI is significantly reduced.

Code frames for some kinds of data cannot be set up prior to data collection either because the kinds and amounts of data that may be obtained cannot be anticipated, or exhaustive, mutually exclusive codes that can be easily explained to and used by data collectors and data entry personnel cannot be created. In such cases, code frames are created inductively after data collection is completed. The process by which this is done is complex and beyond the scope of this entry.

—Linda B. Bourque

REFERENCE

- Bourque, L. B., & Clark, V. A. (1992). *Processing data: The survey example* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-085). Newbury Park, CA: Sage.

CODING QUALITATIVE DATA

CODING is an essential part of the many types of social research. Once data collection has been conducted, the researcher will need to begin to make sense of the data. Coding requires the researcher to interact with and to think about the data. It is a systematic way in which to condense extensive data sets into smaller analyzable units through the creation of categories and concepts derived from the data. Coding makes links between different parts of the data that are regarded as having common properties. Coding facilitates the

organization, retrieval, and interpretation of data and leads to conclusions on the basis of that interpretation.

Although the process of coding may be similar for both quantitative and qualitative data, as they both include the activity of sorting the data, the objectives of coding are rather different. For example, when coding data from a social survey, values will be allocated to a category so that it can be statistically managed (frequency counts may be conducted). By contrast, the categories obtained in a SEMI-STRUCTURED INTERVIEW are not prescribed values; rather, they are explored through themes and remain embedded in their contextual position. Qualitative coding enables the data to be identified, reordered, and considered in different ways.

Whether coding open-ended questions on questionnaires, interview transcripts, or field notes, the main stage in coding is to develop a set of categories into which the data are coded. These categories may be theory driven or data driven, derived from research literature, or based on intuition. Strauss and Corbin (1990) identify three different levels of coding: open, axial, and selective. Open coding involves breaking down, comparing, and categorizing the data; axial coding puts the data back together through making connections between the categories identified after open coding; and selective coding involves selecting the core category, relating it to other categories while confirming and explaining these relationships. When a research project is designed to test a hypothesis, the categories will be generated from the theoretical framework employed, and the data will be molded into these categories. This type of coding is termed *coding down*. When data are used to generate theory, the categories are generated from the data, which involves *coding up* (Fielding, 1993).

Coffey and Atkinson (1996) highlight that coding can serve as either data simplification or data complication, although often coding endorses a combination of the two. Coding can reduce data to simple broad categories or common denominators that help the retrieval of data sharing common features. As demonstrated by Strauss (1987), coding can equally be employed to expand on, reinterpret, and open up analysis to previously unconsidered analytic possibilities and aid the generation of theories.

Over the past decade, there has been an explosion of computer software packages specifically designed for facilitating the qualitative coding process. Although computational qualitative coding can be conducted quickly and simply, be used with large numbers of

coding, and open up unexpected lines of enquiry and analysis, it can offer the researcher an overwhelming range of ways in which textual data may be handled. Furthermore, research can be constrained by the computer program. It may impose unacceptable restrictions on analysis, and qualitative coding can be subject to formalization.

—Sharon Lockyer

REFERENCES

- Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data: Complementary research strategies*. Thousand Oaks, CA: Sage.
- Fielding, J. (1993). Coding and managing data. In N. Gilbert (Ed.), *Researching social life* (pp. 218–238). Thousand Oaks, CA: Sage.
- Strauss, A. (1987). *Qualitative analysis for social scientists*. Cambridge, UK: Cambridge University Press.
- Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.

COEFFICIENT

A coefficient is the quantitative value of a RELATIONSHIP. A coefficient, once estimated, indicates the link between two or more variables. For example, a PEARSON CORRELATION COEFFICIENT between variables *X* and *Y* of 0.90 precisely quantifies that relationship, suggesting it is very strong. In a MULTIPLE REGRESSION equation, with two independent variables *X* and *Z*, a regression coefficient of 6.3 for *X* indicates that a unit change in *X* is expected to produce a 6.3-unit change in *Y*, holding *Z* constant. In general, there are many coefficients available for relating variables, with each having different and, in some cases, more desirable properties, depending on the purposes of the analyst.

—Michael S. Lewis-Beck

See also PARAMETER

COEFFICIENT OF DETERMINATION

Represented by r^2 for the bivariate case and R^2 in the multivariate case, the coefficient of determination is a measure of GOODNESS OF FIT in ORDINARY LEAST

SQUARES LINEAR REGRESSION. (In the multivariate case, it is often called the coefficient of multiple determination.) The statistic measures how well the estimated regression line fits the actual data. It tells us how much of the variation in the dependent variable, *y*, is explained by the model by all of the INDEPENDENT VARIABLES taken together. Specifically, it is a proportion of the explained variation in *y* to the total variation in *y*.

A number of equations may be used to calculate the coefficient of determination. The first takes the explained sum of squares and divides it by the total sum of squares: $R^2 = \text{ESS}/\text{TSS}$. A second subtracts the proportion of the residual sum of squares (the unexplained sum of squares) to the total sum of squares from 1: $R^2 = 1 - (\text{RSS}/\text{TSS})$.

Because it is the proportion of the explained sum of squares over the total sum of squares, the measure must fall between 0 and 1. A value of 1 indicates a perfect fit of the linear regression line to the data; values closer to 0 suggest a poor fit. If the *x* and *y* variables are completely linearly independent, the R^2 will equal 0. It should be noted that R^2 could be a negative value ranging from 0 to –1 if the sample average accounts for more variation in the dependent variables than the independent variables explain. Multiplying the measure by 100 gives a value that allows for clearer interpretation; with this calculation, then, an R^2 of .25 would be interpreted by saying that 25% of the variation in the dependent variable is explained by the independent variables in the model.

The square root of the coefficient of determination is known as the coefficient of MULTIPLE CORRELATION, or the sample CORRELATION coefficient in the bivariate case. Conversely, R^2 is the square of the coefficient of multiple correlation, which is the measure of correlation between the estimated dependent variable calculated from the independent variables ($\hat{y}$) and the actual dependent variable (*y*).

Although much consideration, perhaps sometimes too much, is often paid to this measure, there are a number of important points to remember when interpreting it. First, no matter how high your R^2 is, it only is evidence of correlation; it does not provide positive support for causation. That is, a high R^2 does not allow you to state that your independent variables caused your dependent variable. Second, although, for example, 30% is not an extremely high value, your model may be performing relatively well; explaining 30% of the variation of a factor in the political environment is

often still a substantial portion. Also, when conducting TIME-SERIES analysis, you may often find R^2 s over .90. Third, as independent variables are added to the model, the R^2 will increase, but we should avoid trying to maximize the coefficient of correlation over theoretically sound MODELS. Finally, R^2 s cannot be compared between any two given models if they do not have the same dependent variable.

When working with models with multiple variables, it is good practice to consider the ADJUSTED R^2 . Because every independent variable added to a model usually increases the amount of variation explained, it is important to take into consideration the DEGREES OF FREEDOM. The adjusted R^2 does just this by adjusting for the additional parameters.

—Christine Mahoney

REFERENCES

- Frankfort-Nachmias, C., & Nachmias, D. (2000). *Research methods in the social sciences* (6th ed.). New York: Worth.
- Gujarati, D. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.
- Wooldridge, J. M. (2000). *Introductory economics: A modern approach*. Cincinnati, OH: South-Western College Publishing.

COGNITIVE ANTHROPOLOGY

Cognitive anthropology is the study of human thought in the context of social relations. Some work in cognitive anthropology examines discrete cognitive processes such as memory or perception, but Goodenough (1957) set the stage for the modern development of cognitive anthropology when he defined culture as the knowledge an individual must possess to function adequately as a member of a social group. Implicit in this deceptively simple definition are questions regarding both the sharing and distribution of knowledge that are now at the core of the field.

Cognitive anthropology has been one of the most methodologically self-conscious topical areas in anthropology, driven in part by these questions of sharing and distribution. Much of this work revolves around the elicitation of culturally salient categories within a social group (be it a social network, an organization, or a community) and the determination of underlying dimensions of meaning that organize those categories. The techniques for eliciting terms that respondents use

to label categories and for determining what distinctive features respondents use to define the meaning of terms, include the following: free-listing, in which terms that make up a cultural domain are generated; various forms of unconstrained and constrained pile sorts, in which the similarities and differences in the meaning of terms can be determined; and various forms of rating and ranking tasks to test in a refined way the importance of different dimensions of meaning (see Weller & Romney, 1988).

The examination of the sharing and distribution of cultural knowledge advanced dramatically with the development of the cultural consensus model by Romney, Weller, and Batchelder (1986). The cultural consensus model uses the correlations among subjects as a measure of agreement; if these associations among respondents can be accounted for by a single principal factor, then consensus exists, and it is reasonable to infer that all subjects are relying on a single, shared cultural model of that particular cultural domain. The distribution of subjects around that consensus can be examined (including the potential for alternative cultural models), as well as the specific content of the consensus.

A good example of this approach is the study of cultural models of reproductive cancers among physicians and patients from different ethnic backgrounds (Chavez, Hubbel, McMullin, Martinez, & Mishra, 1995). The tools described above were used to examine models of cancer among physicians, Anglo women, and both resident and immigrant Latinas in Southern California. Within each of these groups, there was significant consensus (i.e., groups agreed among themselves about the important causes of reproductive cancers). There was also an overlapping consensus that connected the Anglo women to the physicians, Anglo women to resident Latinas, and resident Latinas to immigrant Latinas. But immigrant Latinas were far from both Anglo women and physicians in their shared models of cancer causation. These results can be displayed graphically, demonstrating how persons and groups are distributed in a space of cultural meaning.

The systematic ethnographic techniques of cognitive anthropology enhance the traditional qualitative techniques of cultural anthropology. A primary goal of anthropological research has been to understand how other people see their worlds. Methodological innovations in both the collection and analysis of data

are helping investigators to understand others' cultural models with greater precision.

—William W. Dressler

REFERENCES

- Chavez, L., Hubbell, F. A., McMullin, J. M., Martinez, R. G., & Mishra, S. I. (1995). Structure and meaning in models of breast and cervical cancer risk factors: A comparison of perceptions among Latinas, Anglo women and physicians. *Medical Anthropology Quarterly*, 9, 40–74.
- Goodenough, W. (1957). Cultural anthropology and linguistics. In P. L. Garvin (Ed.), *Report of the 7th Annual Roundtable on Linguistics and Language Study* (pp. 167–173). Washington, DC: Georgetown University Press.
- Romney, A. K., Weller, S. C., & Batchelder, W. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, 88, 313–338.
- Weller, S. C., & Romney, A. K. (1988). *Systematic data collection* (Qualitative Research Methods, Vol. 10). Newbury Park, CA: Sage.

COHORT ANALYSIS

Cohort analysis, a general strategy for examining data rather than a statistical technique, has become increasingly popular in the social sciences in the past few decades, especially during the last quarter of the 20th century. Useful in assessing the consequences of aging (of humans or other entities) and in understanding social and cultural change, cohort analysis has also attracted attention because it presents an unusually intriguing methodological challenge.

In its broadest sense, cohort analysis is any quantitative research that uses a measure of the concept of *cohort* and relates that measure to one or more additional variables. A cohort, in turn, consists of those individuals (human or otherwise) who experienced a particular event during a specified time. The kind of cohort most often studied is the human birth cohort—those persons born during a given year, decade, or other period of time. If the term *cohort* is used without a modifier, it is usually understood that the referent is a human birth cohort. Otherwise, the event commonly experienced by the individuals is used as an adjective to identify the kind of cohort, as in *retirement cohort*. The events that define cohorts may range from marriage to joining an organization, from entering a graduate program to becoming a parent for the

first time. The individuals, if not humans, may be marriages, organizations, textbooks, movies, or other entities that came into being during a particular time.

The usual and more restricted meaning of cohort analysis is that it is any attempt to estimate age, period, and cohort effects on a dependent variable, such as some kind of behavior or attitudes, and such attempts require data from at least two cohorts for at least two points in time. An age effect results from growing older, a cohort effect from cohort membership, and a period effect from influences that vary through time. For example, an enduring preference for the kind of music that was popular when one was a teenager or young adult is a cohort effect, whereas changes in attitudes in response to an economic downturn are PERIOD EFFECTS. It is obviously important to estimate age effects in studies of human aging, and it is also important to estimate each kind of effect to understand social and cultural change.

THE AGE-PERIOD-COHORT CONUNDRUM

Estimation of age, period, and cohort effects cannot be straightforward because two of the kinds of effects are confounded with one another in the findings from any standard statistical analysis. For instance, differences between people of different ages at one point in time may be age effects, cohort effects, or both kinds of effects. Likewise, observed changes in persons studied in different waves of a panel study may be age effects, period effects, or both. This confounding of effects is an example of the IDENTIFICATION PROBLEM, which exists when three or more independent variables may affect a dependent variable when each independent variable is a perfect linear function of the others. Age is a linear function of period and cohort, cohort is a linear function of age and period, and period is a linear function of age and cohort.

A heuristic device often used to facilitate understanding of the confounding of effects in cohort data is the standard cohort table, in which two or more sets of cross-sectional data relating some dependent variable to age are juxtaposed and the age categories are equal in years to the intervals between the sets of data, as in Table 1. Period and cohort effects are confounded in the rows of the table, and age and cohort effects are confounded in the columns. Cohorts can be traced through time by reading diagonally down and to the right, and age and period effects are confounded in the cohort diagonals.

Table 1 Percentage of Respondents to the 1974, 1984, and 1994 American General Social Surveys Who Said That Extramarital Sex Relations Are “Always Wrong,” by Age

	1974	1984	1994
Age	% (n)	% (n)	% (n)
20–29	59.2 (392)	68.3 (384)	83.8 (346)
30–39	70.9 (291)	63.9 (304)	77.7 (462)
40–49	75.3 (270)	70.8 (241)	76.7 (411)
50–59	80.8 (278)	78.6 (181)	77.2 (272)
60–69	83.0 (194)	79.7 (166)	85.8 (165)

NOTE: Data are weighted by the number of respondents in the household age 18 and older to increase representativeness.

The linear effects of age, period, and cohort are confounded with one another in such a way that no statistical analysis can separate them; an infinite number of combinations of age, period, and cohort effects could account for a linear pattern of variation of a dependent variable across ages, periods, or cohorts. There is a method that can separate nonlinear effects, but only if the effects are additive and if precisely correct assumptions can be made about some of the effects—conditions that are rarely met.

However, it is often possible to arrive at reasonable conclusions about the general nature of the effects without rigorous model testing. Some useful cohort analyses have combined theory, information about the phenomena being studied from outside the data at hand, common sense, simple statistical analyses, and informal examination of the data to arrive at tentative but defensible conclusions.

AN EXAMPLE

The utility of a simple examination of descriptive data can be illustrated by looking at the data in Table 1. The 1974 data show a positive relationship between age and restrictive attitudes concerning extramarital sex relations, and the cohort diagonals indicate that, in the aggregate, persons in each birth cohort grew more restrictive as they grew older. These data by themselves suggest a positive age effect on restrictive attitudes. However, a more careful examination of the table casts doubt on this interpretation. Restrictiveness increased at the 20 to 29 age level—an indication of period influences toward restrictive attitudes possibly

strong enough to produce the increased restrictiveness within cohorts. According to theory, supported by much empirical evidence, younger adults tend to be affected more than older ones by period influences; thus, such a differential impact by age of earlier influences for permissiveness could account for the positive relationship between age and restrictiveness in 1974. Therefore, there is no convincing evidence of any age effects in the data.

AGE-PERIOD-COHORT CHARACTERISTIC MODELS

A new form of analysis, introduced by Robert O’Brien and his collaborators, can reasonably be considered a kind of cohort analysis, even though it does not provide meaningful estimates of age, period, or total cohort effects. Although it is impossible to hold constant two of the interrelated variables and let the third vary, it is possible to hold age and period constant and let certain characteristics of cohorts, such as their relative size, vary. If the cohort characteristic does not bear a strong linear relationship to cohort, the estimate of its effects should be meaningful, even though the model estimates of age and period effects are not. This promising technique has already proven useful in studying such dependent variables as crime commission and victimization.

—Norval D. Glenn

REFERENCES

- Glenn, N. D. (2003). Distinguishing age, period, and cohort effects. In J. Mortimer & M. Shanahan (Eds.), *Handbook of the life course* (pp. 465–476). New York: Kluwer Academic/Plenum.
- Mason, K. O., Mason, W. M., Winsborough, H. H., & Poole, W. K. (1973). Some methodological issues in the cohort analysis of archival data. *American Sociological Review*, 38, 242–258.
- O’Brien, R. M. (2000). Age-period-cohort-characteristic models. *Social Science Research*, 29, 123–139.

COINTEGRATION

When a linear combination of nonstationary variables is stationary, the variables are said to be cointegrated, and the vector that defines the stationary linear combination is called a *cointegration vector*. A

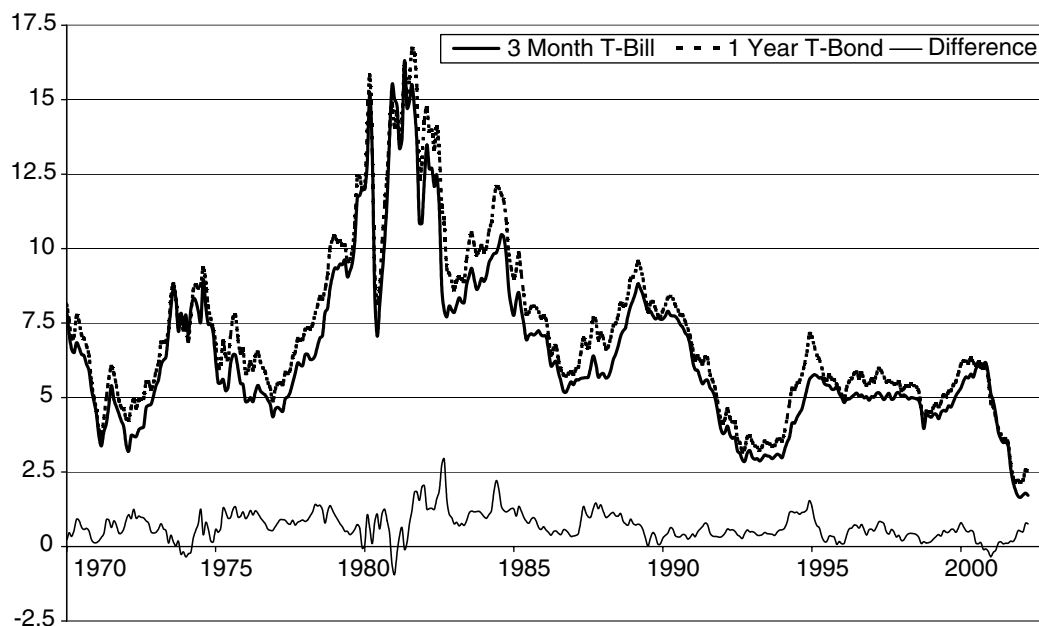


Figure 1 The Interest Rates on a 3-Month Treasury Bill and a 1-Year Treasury Bond for the Period From January 1970 to April 2002

SOURCE: Federal Reserve Bank of St. Louis.

NOTE: The two interest rates appear to be nonstationary, whereas the difference between them seems to be stationary, so the two interest rates are cointegrated.

time series is stationary if its distribution does not vary over time. The simplest example of a stationary process is $\{\varepsilon_t\} = \dots, \varepsilon_{-1}, \varepsilon_0, \varepsilon_1, \dots$, which represents a sequence of independent and identically distributed random variables. The subscript t refers to time. If the distribution of a variable depends on t , it is nonstationary. In cointegration analysis, the most common form of nonstationarity is that of the integrated variables. The random walk, $X_t = X_{t-1} + \varepsilon_t = X_0 + \sum_{i=1}^t \varepsilon_i$, is an example of a nonstationary variable that is integrated of order one. The word *integration* refers to the cumulation of epsilons.

The concept of cointegration is interesting because it can be applied to uncover relationships that may have theoretical interpretations. For example, the cointegration relations may be defined by the first-order conditions of an economic model, and cointegration analysis can be used to estimate economic models and test certain theoretical hypotheses.

The classical example on cointegration is about a dog that follows its drunk owner. In this example, the positions of the dog and its owner, as a function of time, are two nonstationary processes because the drunk owner is walking around, taking steps in

random directions. However, the two processes are cointegrated because the distance between dog and owner is stationary.

Another example is a property of interest rates on bonds with different maturities. Individually, they may have large fluctuations, but the difference between them appears to be stationary. This is illustrated in Figure 1, which contains the interest rates on a 3-month Treasury bill and a 1-year Treasury bond for the period from January 1970 to April 2002. The thin solid line is the difference between the two interest rates.

HISTORICAL DEVELOPMENT

Cointegration was introduced by Clive W. J. Granger (1981, 1983), and the statistical analysis of cointegrated processes was formalized by Robert F. Engle and Granger (1987), Søren Johansen (1988, 1991), and Peter C. B. Phillips (1991). Cointegration is related to many concepts of *ECONOMETRICS*, such as unit root processes, spurious regression, and common stochastic trends (for an excellent review, see Watson, 1994). Today there is a voluminous literature on cointegration and applications of

cointegration. Recent developments have been on improving and generalizing existing techniques, such as bias and Bartlett corrections; fractionally cointegrated processes; seasonal cointegration; panel cointegration; nonlinear cointegration; and cointegration in relation to processes with structural changes. Many references can be found in the book by James D. Hamilton (1994), which also contains a good introduction to cointegration. Excellent textbooks on likelihood analysis of cointegration are Johansen (1995) and the companion book by Peter R. Hansen and Johansen (1998).

ORDER OF INTEGRATION

For variables to be cointegrated, they must individually be integrated. Formally, a time series, X_t , is said to be integrated of order d if $(1 - L)^d X_t = \psi(L)\varepsilon_t$ is stationary and $\psi(1) = \sum i \psi_i \neq 0$, where L is the lag operator such that $\psi(L)\varepsilon_t = \psi_0 \varepsilon_t + \psi_1 \varepsilon_{t-1} + \dots$. The notation $X_t \sim I(d)$ is short for “ X_t is integrated of order d ,” and the notation $\Delta^d X_t$ is sometimes used in place of $(1 - L)^d X_t$. An alternative definition of the order of integration assumes that $\{\varepsilon_t\}$ is a white-noise process and substitutes *covariance-stationary* for *stationary*. A white-noise process is defined by the following: having expected value zero, $E(\varepsilon_t) = 0$; having finite variance that does not depend on time, $\text{var}(\varepsilon_t) = \sigma^2$; and being uncorrelated across time, $\text{cov}(\varepsilon_s, \varepsilon_t) = 0$ for $s \neq t$. A covariance-stationary process, $\{X_t\}$, has constant mean and variance, as well as an autocovariance function that only depends on the difference in time, $\text{cov}(X_s, X_t) = \varphi(|s - t|)$.

The order of integration, d , can be any real number, but most of the literature is concerned with the unit root processes, which corresponds to $d \in \mathbb{N}$ (the positive integers). The nonintegers are referred to as fractionally integrated processes, which is related to the long-memory processes.

The order of integration of a vector of variables, $X_t = (X_{1t}, \dots, X_{pt})'$, is defined as the highest order of integration of the individual elements, $X_{1t}, \dots, X_{pt}$.

If $X_t \sim I(d_x)$, if $Y_t \sim I(d_y)$, and if a and b are nonzero constants, then $aX_t + bY_t$ is, in general, integrated of order $d = \max(d_x, d_y)$. It is the special case when $aX_t + bY_t$ is integrated of order less than $\max(d_x, d_y)$, which characterizes cointegration. For two variables to be cointegrated, they must be integrated of the same order. The notation $CI(d, b)$ is used for variables that are integrated of order d

with a cointegration relation, which is integrated of order $d - b$.

A more general concept of cointegration is multi-cointegration, which is sometimes called polynomial cointegration. If $(1 - L)^d X_t$ and Y_t are cointegrated for some $d \neq 0$, then X_t and Y_t are multi-cointegrated. For example, $X_t \sim I(2)$ and $Y_t \sim I(1)$, then $\Delta X_t = X_t - X_{t-1}$ may cointegrate with Y_t .

EXAMPLE

Let $\{\varepsilon_t\}$ be a sequence of independent and identically distributed variables, and suppose that for $t = 1, 2, \dots$

$$X_t = \sum_{s=1}^t \sum_{i=1}^s \varepsilon_i + \sum_{s=1}^t \varepsilon_s + \varepsilon_t,$$

$$Y_t = 2 \sum_{s=1}^t \sum_{i=1}^s \varepsilon_i + \varepsilon_t,$$

$$Z_t = - \sum_{s=1}^t \varepsilon_s + \varepsilon_t.$$

Because $(1 - L)^2 X_t = 3\varepsilon_t - 3\varepsilon_{t-1} + \varepsilon_{t-2}$ is stationary and the coefficients satisfy $\sum_i \psi_i = 3 - 3 + 1 = 1 \neq 0$, we see that X_t is integrated of order two. Similarly, it follows that $Y_t \sim I(2)$ and that $Z_t \sim I(1)$, so the vector $(X_t, Y_t, Z_t)'$ is integrated of order two. Furthermore, $2X_t - Y_t = 2 \sum_{s=1}^t \varepsilon_s + \varepsilon_t$ is integrated of order one, so X_t and Y_t are cointegrated, $CI(2, 1)$, with cointegration vector $(2, -1)'$. Similarly, we see that $2X_t - Y_t + 2Z_t = 3\varepsilon_t$ is integrated of order zero, so the three variables— X_t , Y_t , and Z_t —are cointegrated, $CI(2, 2)$, with cointegration vector $(2, -1, 2)'$. Also, X_t and Z_t are multi-cointegrated because the two $I(1)$ variables, ΔX_t and Z_t , are cointegrated, $CI(1, 1)$, because $\Delta X_t + Z_t = 3\varepsilon_t - \varepsilon_{t-1}$ is $I(0)$. Similarly, it can be seen that Y_t and Z_t are multi-cointegrated.

ANALYSIS OF COINTEGRATED PROCESSES

The classical assumptions in the LINEAR REGRESSION model are violated in the presence of integrated variables, mainly because of the nonstationarity. This can cause the asymptotic normality of parameter estimators to fail, and the statistical analysis of cointegrated variables is therefore more complicated than standard regression analysis.

The two main approaches to cointegration analysis are the regression-based approach and the likelihood approach. The former involves standard regression techniques with various modifications, and the latter, which is based on the vector autoregressive model (VAR), is sometimes called the Johansen method.

A cointegration analysis of a system of variables will typically consist of the following:

1. *Determining the cointegration rank.* This typically involves the use of a Dickey-Fuller-type distribution. These distributions have complicated expressions that involve stochastic integrals of Brownian motions. Fortunately, most modern econometrics textbooks, including the one by Hamilton (1994), contain tables of the most common Dickey-Fuller distributions. The trace test of Johansen (1988, 1991) is a popular test for determining the cointegration rank.

2. *Estimating cointegration relations and other parameters.* The cointegration relations are not identified without additional restrictions/normalizations.

3. *Inference on the cointegration parameters and other parameters.* The analysis on cointegration relations is particularly useful because they may be interpreted as structural equations in an economic model. Identified cointegration parameters have a nonnormal limiting distribution and are extremely consistent. The latter means that small samples can be very informative about the cointegration relations. Inference on other parameters follows traditional results in the standard case.

The likelihood analysis of Johansen (1988) is based on the VAR model, $X_t = \prod_1 X_{t-1} + \cdots + \prod_k X_{t-k} + \varepsilon_t$, where X_t is a p -dimensional vector and $\{\varepsilon_t\}$ is a sequence of independent normally distributed variables with mean 0 and covariance-matrix Ω . This model is rewritten in the form of the error correction model (ECM):

$$\begin{aligned} \Delta X_t &= \prod_1 X_{t-1} + \Gamma_1 \Delta X_{t-1} \\ &\quad + \cdots + \Gamma_{k-1} \Delta X_{t-k+1} + \varepsilon_t, \\ t &= 1, \dots, T. \end{aligned}$$

If the elements of X_t are cointegrated, then the $p \times p$ matrix, $\prod$, has reduced rank $r < p$, and the number of common stochastic trends in X_t is given by $p - r$. The reduced rank of $\prod$ can be made explicit by expressing the matrix as a product of two matrices, $\prod = \alpha\beta'$, where α and β are $p \times r$ matrices. This formulation leads to a regression equation that can be estimated by reduced rank regression. An important instrument for

the analysis of cointegrated processes is the Granger representation, $X_t = C \sum_{i=1}^t \varepsilon_i + C(L)\varepsilon_t + X_0 - C(L)\varepsilon_0$. This representation divides the process into a nonstationary random walk, $C \sum_{i=1}^t \varepsilon_i$; a stationary component, $C(L)\varepsilon_t$; and a term that depends on initial values, $X_0 - C(L)\varepsilon_0$. An important result is that $\beta'C = 0$ (an $r \times p$ matrix of zeros), from which it follows that $\beta'X_t = \beta'C(L) \sim I(0)$. So the cointegration relations are given by the columns of β . The relation, $\Delta X_t = \alpha(\beta'X_{t-1}) + \cdots$, shows that α defines how X_t responds on a deviation in $\beta'X_{t-1}$ from its expected value, and α is therefore called the matrix of adjustment coefficients. $\beta'X_{t-1}$ can sometimes be interpreted as the deviation from an equilibrium. In this case, α informs about the variables' adjustment toward the equilibrium relations, both the direction they take and how fast the variables converge to the equilibrium.

—Peter Reinhard Hansen

REFERENCES

- Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica*, 55, 251–276.
- Granger, C. W. J. (1981). Some properties of time series data and their use in econometric models specification. *Journal of Econometrics*, 16, 121–130.
- Granger, C. W. J. (1983). *Co-integrated variables and error-correcting models*. Discussion Paper No. 1983–13, University of California, San Diego.
- Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.
- Hansen, P. R., & Johansen, S. (1998). *Workbook on cointegration*. Oxford, UK: Oxford University Press.
- Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12, 231–254.
- Johansen, S. (1991). Estimation and hypotheses testing of cointegrating vectors in Gaussian vector autoregressive models. *Econometrica*, 59, 1551–1580.
- Johansen, S. (1995). *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford, UK: Oxford University Press.
- Phillips, P. C. B. (1991). Optimal inference in cointegrated systems. *Econometrica*, 59, 283–306.
- Watson, M. W. (1994). Vector autoregression and cointegration. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of econometrics* (Vol. 4, pp. 2843–2915). Amsterdam: Elsevier.

COLLINEARITY. See MULTICOLLINEARITY

COLUMN ASSOCIATION. See ASSOCIATION MODEL

COMMONALITY ANALYSIS

Commonality analysis is a method used to partition the explained/predicted variance in either a measured or a latent variable into subcomponents explained (a) uniquely by each measured predictor variable and (b) in common by every possible combination of the PREDICTOR VARIABLES. The explained variance and all the subcomponents of the explained variance are each in a squared, variance-accounted-for metric. Commonality analysis was originally developed for use in REGRESSION but also can be applied in other MULTIVARIATE ANALYSES (cf. Frederick, 1999).

HEURISTIC EXAMPLE

To make the discussion concrete, presume a regression problem involving a single measured outcome variable (Y) and two predictor variables (X_1 and X_2). The results associated with the example are presented graphically in the Venn diagram in Figure 1.

In this example, each of the three measured variables has a SUM OF SQUARES (SOS) of 400, corresponding to the box for each measured variable being drawn as encompassing 20×20 units of area. As regards the overlap of X_1 and Y , because 100 units of the $SOS_Y = 400$ are predicted by X_1 , the $r^2_{YX_1} = 100/400 = .2500 = 25.00\%$. X_2 explains 225 of the 400 units of Y , and so the $r^2_{YX_2} = 225/400 = .5625 = 56.25\%$.

However, the MULTIPLE CORRELATION squared ($R^2_{[YX_1X_2]}$) does *not* equal 25.00% plus 56.25%, or 81.25%, because the predictors are correlated ($r^2_{X_1X_2} = 18.75\%$). Therefore, some of the explanatory power of X_1 and X_2 is common to both and should not be “double counted” in computing $R^2_{[YX_1X_2]}$. For this example, X_1 and X_2 together explain $R^2_{[YX_1X_2]} = 68.75\%$ of the variability in the Y scores (and not 81.25%).

But where does this 68.75% EFFECT SIZE originate? The following questions must be addressed:

1. How much of the observed effect size is due to the *unique* explanatory power of X_1 ?

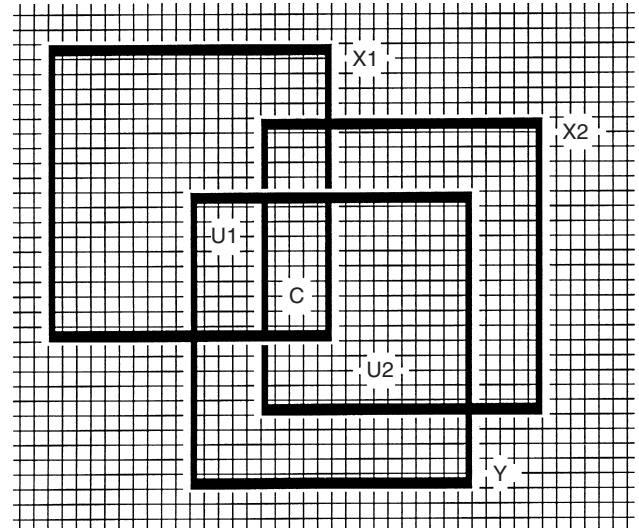


Figure 1 Venn Diagram for a Two-Predictor Regression Commonality Analysis

2. How much of the observed effect size is due to the *unique* explanatory power of X_2 ?
3. How much of the observed effect size is due to the *common* explanatory power of X_1 and X_2 ?

These are the questions addressed in a commonality analysis.

The analysis is conducted in a series of four steps. First, the R^2 is computed using all possible combinations of the predictor variables. For this simple example, there are only three possible combinations, and $R^2_{[YX_1X_2]} = 68.75\%$, $R^2_{[YX_1]} = 25.00\%$, and $R^2_{[YX_2]} = 56.25\%$.

Second, values obtained in Step 1 are inserted into commonality analysis formulas provided in various publications (cf. Rowell, 1996). For this example, the unique predictive contribution (U_1) of X_1 is computed as

$$U_1 = R^2_{[YX_1X_2]} - R^2_{[YX_2]}$$

$$68.75\% - 56.25\% = 12.50\%.$$

The unique contribution (U_2) of X_2 is computed as

$$U_2 = R^2_{[YX_1X_2]} - R^2_{[YX_1]}$$

$$68.75\% - 25.00\% = 43.75\%.$$

The common contribution (C_{12}) of X_1 and X_2 is computed as

$$C_{12} = R^2_{[YX_1X_2]} - U_1 - U_2$$

$$68.75\% - 12.50\% - 43.75\% = 12.50\%.$$

The commonality analysis computational formulas become exponentially more complicated as there are more predictor variables (Rowell, 1996). For example, when there are three predictor variables, the formulas are as follows:

$$U_1 = R^2_{[YX_1X_2X_3]} - R^2_{[YX_2X_3]}$$

$$U_2 = R^2_{[YX_1X_2X_3]} - R^2_{[YX_1X_3]}$$

$$U_3 = R^2_{[YX_1X_2X_3]} - R^2_{[YX_1X_2]}$$

$$C_{12} = R^2_{[YX_1X_3]} + R^2_{[YX_2X_3]} - R^2_{[YX_3]} - R^2_{[YX_1X_2X_3]}$$

$$C_{13} = R^2_{[YX_1X_2]} + R^2_{[YX_2X_3]} - R^2_{[YX_2]} - R^2_{[YX_1X_2X_3]}$$

$$C_{23} = R^2_{[YX_1X_2]} + R^2_{[YX_1X_3]} - R^2_{[YX_1]} - R^2_{[YX_1X_2X_3]}$$

$$C_{123} = R^2_{[YX_1X_2X_3]} + R^2_{[YX_1]} + R^2_{[YX_2]} + R^2_{[YX_3]}$$

$$- R^2_{[YX_1X_2]} + R^2_{[YX_1X_3]} - R^2_{[YX_2X_3]}.$$

Third, as a check on the computations, the $R^2_{[YX_1X_2]}$ is recomputed as

$$R^2_{[YX_1X_2]} = U_1 + U_2 + C_{12}$$

$$12.50\% + 43.75\% + 12.50\% = 68.75\%.$$

Note, as illustrated in Figure 1, that $U_1 = 12.50\%$ and $C_{12} = 12.50\%$ are *not* the same areas.

Fourth, the variance decompositions are then organized into a table to illustrate the unique and common predictive contributions of each predictor variable individually. Table 1 presents the relevant results for this example.

COMMON MISCONCEPTION

A common misconception of persons using REGRESSION is that the predictor variables should be uncorrelated (i.e., not COLLINEAR or MULTICOLLINEAR). It is true that when predictors are correlated, sensible interpretations as regards the origins of detected EFFECT SIZES *cannot* be derived by interpreting only standardized weights ("beta weights"). It is also true that standard errors of the standardized weights become larger when predictors are correlated.

But CORRELATIONS among predictor variables are not inherently problematic as long as correct interpretation strategies are employed. Using correlated

Table 1 Variance Partitions of the Squared Bivariate Correlations (in percentages)

Variance Partition	Predictor	
	X_1	X_2
U_1	12.50	
U_2		43.75
C_{12}	12.50	12.50
Sum = r^2	25.00	56.25

predictors also often honors the fact that, in reality, many variables are correlated. One such interpretation strategy invokes examination of both the standardized weights and the structure coefficients (Courville & Thompson, 2001).

Commonality analysis is another alternative for understanding prediction dynamics when predictors are correlated. Note that when predictor variables are perfectly uncorrelated, a commonality analysis would not be sensible because every common variance partition (e.g., C_{12} , C_{23} , C_{123}) would, by definition, be zero, and every unique contribution (e.g., U_1) would equal a corresponding squared bivariate correlation (e.g., $r^2_{YX_1}$).

NEGATIVE VARIANCE COMPONENTS

In any analysis, the results are compromised if there is excessive model specification error. So, if the wrong variables or the wrong analysis is used in a commonality analysis, the results are compromised. One indication of this possibility occurs when some of the variance components are negative because squared variance-accounted-for statistics should not be negative (Thompson, 1985).

If one or a few variance components in a commonality analysis are negative but only trivially smaller than zero, the negative values are only mildly distressing. However, if the negative values are large in magnitude in the judgment of the researcher, careful thought must be invested in whether the model has been correctly specified.

—Bruce Thompson

REFERENCES

- Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is

not enough. *Educational and Psychological Measurement*, 61, 229–248.

Frederick, B. N. (1999). Partitioning variance in the multivariate case: A step-by-step guide to canonical commonality analysis. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 305–318). Stamford, CT: JAI.

Rowell, R. K. (1996). Partitioning predicted variance into constituent parts: How to conduct commonality analysis. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 4, pp. 33–44). Greenwich, CT: JAI.

Thompson, B. (1985). Alternate methods for analyzing data from experiments. *Journal of Experimental Education*, 54, 50–55.

COMMUNALITY

Communality is a squared variance-accounted-for statistic reflecting how much variance in measured variables is reproduced by the latent constructs (e.g., the factors) in a model. Conversely, communality can be conceptualized as how much of the variance of a measured/observed variable is useful in delineating the latent/composite variables in the model. The symbol typically used for a communality coefficient is h^2 .

Communality coefficients are commonly used in FACTOR ANALYSIS, including both EXPLORATORY FACTOR ANALYSIS (EFA) and CONFIRMATORY FACTOR ANALYSIS (CFA) (Graham, Guthrie, & Thompson, 2003; Thompson, 1997). However, communality coefficients can also be computed in other multivariate techniques, such as CANONICAL CORRELATION ANALYSIS (Thompson, 2000). When variables are being analyzed, there will be a separate communality coefficient computed for each variable.

Communality coefficients are estimated both retrospectively, after an analysis is completed, and prospectively, as input into initial data analysis. In the retrospective application, if the factors are uncorrelated, the communality coefficient for a measured variable is computed by squaring the structure coefficients (factor loadings) for the variable on all factors and then summing these squared structure coefficients across the factors.

For example, 10 variables might be subjected to an EFA in which three factors were extracted and then subjected to ORTHOGONAL ROTATION. The squared structure coefficient for the first variable on the three factors might be .4, .3, and .2. The h^2 for this variable would be .9 or 90% (i.e., $.4 + .3 + .2 = .9$). This means

that 90% of this measured variable's variance can be reproduced by the three factors. Conversely, 90% of the variance of the measured variable was useful in delineating the three factors as a set.

Communality coefficients can also be conceptualized as lower-bound estimates of the minimum RELIABILITY COEFFICIENT of a given variable. If a measured variable had an $h^2 = .9$, then whatever the score reliability of the variable is, the reliability is *at least* .9.

The input into an analysis affects the output communalities. Often, the correlation matrix is analyzed. Whatever entries are initially put on the diagonal of the matrix are the input communality estimates. Then factors are extracted. In PRINCIPAL COMPONENTS ANALYSIS, the communality coefficients are computed at this point, and the analysis is complete.

However, in analyses such as principal axis factor analysis, these communality coefficients are then substituted back into the diagonal of the correlation matrix, and factors are reextracted. This process is repeated successively (called iteration) until the communality coefficients on successive steps stabilize.

The influence of iteration on the output results is a function of (a) the score reliabilities of the variables and (b) the number of entries on the diagonal versus the number of entries in the entire matrix (e.g., for 10 variables, $10/100 = .1$; for 100 variables, $100/10,000 = .01$). It should also be noted that what is being estimated is reliability so as to take into account measurement ERROR, but excessive iteration may let SAMPLING ERROR distort the results too much. So some methodologists suggest no or limited iteration to balance the trade-offs between these two types of error influences.

—Bruce Thompson

See also FACTOR ANALYSIS

REFERENCES

- Graham, J. M., Guthrie, A. C., & Thompson, B. (2003). Consequences of not interpreting structure coefficients in published CFA research: A reminder. *Structural Equation Modeling*, 10, 142–153.
- Thompson, B. (1997). The importance of structure coefficients in structural equation modeling confirmatory factor analysis. *Educational and Psychological Measurement*, 57, 5–19.
- Thompson, B. (2000). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 285–316). Washington, DC: American Psychological Association.

COMPARATIVE METHOD

The comparative method is fundamentally a case-oriented, small-*N* technique. It is typically used when researchers have substantial knowledge of each case included in an investigation and there is a relatively small number of such cases, often of necessity. The best way to grasp the essential features of the comparative method is to examine it in light of the contrasts it offers with textbook presentations of social science methodology. In textbook presentations, (a) the proximate goal of social research is to document general patterns that hold for a large population of observations, (b) cases and populations are typically seen as given, (c) researchers are encouraged to enlarge their number of cases whenever possible, (d) it is often presumed that researchers have well-defined theories and well-formulated hypotheses at their disposal from the very outset of their research, (e) investigators are advised to direct their focus on VARIABLES that display a healthy range of variation, and (f) researchers are instructed to assess the relative importance of competing independent variables to understand causation.

PROXIMATE GOALS

Textbook presentations of social research usually focus on the goal of documenting general patterns characterizing a large population of observations. If researchers can demonstrate a relationship between one or more independent variables and a DEPENDENT VARIABLE, then they can better predict cases' values on the dependent variable, given knowledge of their scores on the predictor variables. Typically, the study of general patterns is conducted with a sample of observations drawn from a large population. The researcher draws inferences about the larger population based on his or her analysis of the sample.

By contrast, COMPARATIVE RESEARCH focuses not on relationships between variables or on problems of inference and prediction but on the problem of making sense of a relatively small number of cases, selected because they are substantively or theoretically important in some way (Eckstein, 1975). For example, a researcher might use the comparative method to study a small number of guerilla movements in an in-depth manner. Suppose these movements were all thought to be especially successful in winning popular support. To find out how they did it, the researcher would

conduct in-depth studies of the movements in question. This case-oriented comparative study would differ dramatically from a "textbook" study. In the latter, the researcher would sample guerilla movements from the population of such movements (assuming this could be established) and then characterize these movements in terms of generic variables and their relationships (e.g., the correlation between the extent of foreign financial backing and the degree of popular support).

As these examples show, the key contrast is the researcher's proximate goal: Does the researcher seek to understand specific cases or to document general patterns characterizing a population? This contrast follows a longstanding division in all of science, not just social science. Georg Henrik von Wright (1971, pp. 1–33) argues in *Explanation and Understanding* that there are two main traditions in the history of ideas regarding the conditions an explanation must satisfy to be considered scientifically respectable. One tradition, which he calls "finalistic," is anchored in the problem of making facts understandable. The other, called "causal-mechanistic," is anchored in the problem of prediction. The contrasting goals of comparative research and "textbook" social research closely parallel this fundamental division.

CONSTITUTION OF CASES AND POPULATIONS

Textbook presentations of social science methodology rarely examine the problem of constituting cases and populations. The usual case is the individual survey respondent; the usual population is demarcated by geographic, temporal, and demographic boundaries (e.g., adults in the United States in the year 2003). The key textbook problematic is how to derive a representative sample from the very large population of observations that is presumed to be at the researcher's disposal.

By contrast, comparative researchers see cases as meaningful but complex configurations of events and structures (e.g., guerilla movements), and their cases are typically macrolevel. Furthermore, they treat cases as singular, whole entities purposefully defined and selected, not as homogeneous observations drawn at random from a pool of equally plausible selections. Most comparative studies start with the seemingly simple idea that social phenomena in like settings (such as organizations, countries, regions, cultures, etc.) may parallel each other sufficiently to permit comparing and contrasting them. The clause, "may parallel each

other sufficiently,” is a very important part of this formulation. The comparative researcher’s specification of relevant cases at the start of an investigation is really nothing more than a working hypothesis that the cases initially selected are in fact instances of the same thing (e.g., guerilla movements) and are alike enough to permit comparison. In the course of the research, the investigator may decide otherwise and drop some cases or even whole categories of cases because they do not appear to belong with what seem to be the core cases. Sometimes, this process of sifting through the cases leads to an enlargement of the set of relevant cases and a commensurate broadening of the scope of the guiding concepts (e.g., from guerilla movements to radical opposition movements) or to the development of a typology of cases (e.g., types of guerilla movements).

Usually, this sifting and sorting of cases is carried on in conjunction with concept formation and elaboration (Walton, 1992). Concepts are revised and refined as the boundary of the set of relevant cases is shifted and clarified (Ragin, 1994). Important theoretical distinctions often emerge from this dialogue of ideas and evidence. The researcher’s answers to both “What are my cases?” and “What are these cases of?” may change throughout the course of the research, as the investigator learns more about the phenomenon in question and refines his or her guiding concepts and analytic schemes.

N OF CASES

One key lesson in every course in quantitative social research is that “more cases is better.” More is better in two main ways. First, researchers must meet a threshold number of cases to even apply quantitative methods, usually cited as an N of 30 to 50. Second, the smaller the N , the more the data must conform to the assumptions of statistical methods, for example, the assumption that variables are normally distributed or the assumption that subgroup variances are roughly equal. However, small N s almost guarantee that such assumptions will not be met, especially when the cases are macrolevel. This textbook bias toward large N s dovetails with the implicit assumption that cases are empirically given, not constructed by the researcher, and that they are abundant. The only problem, in this light, is whether the researcher is willing and able to gather data on as many cases as possible, preferably hundreds if not thousands.

By contrast, comparative research is very often defined by its focus on phenomena that are of interest because they are rare—that is, precisely because the N of cases is small. Typically, these phenomena are large scale and historically delimited, not generic in any sense. The population of cases relevant to an investigation is often limited by the historical record to a mere handful. The key contrast with textbook social science derives from the simple fact that many phenomena of interest to social scientists and their audiences are historically or culturally significant. To argue that social scientists should study only cases that are generic and abundant or that can be studied in isolation from their historical and cultural contexts would severely limit both the range and value of social science.

THEORY TESTING

Conventional presentations of social science methodology place great emphasis on theory testing. In fact, its theory-testing orientation is often presented as what makes social science scientific. Researchers are advised to follow the scientific method and develop their hypotheses in isolation from the analysis of empirical evidence. It is assumed further that existing theory is sufficiently well formulated to permit the specification of testable hypotheses and that social scientific knowledge advances primarily through the rejection of theoretical ideas that consistently fail to find empirical support. Of course, the textbook view does not bar from social science those ideas that are generated in a more inductive manner, but any such idea is treated as suspect until it is subjected to an explicit test, using data that are different from those used to generate the idea.

It is without question that theory plays a central role in social research and that, in fact, almost all social research is heavily dependent on theory in some way. However, it is usually not possible to apply the theory-testing paradigm in small- N , case-oriented research. Most comparative research has a very strong inductive component. Researchers typically focus on a small number of cases and may spend many years learning about them, increasing their depth of knowledge. Countless ideas and hypotheses are generated in this process of learning about cases, and even the very definition of the population (What are these cases of?) may change many times in the course of the investigation. Furthermore, because the number of relevant cases in comparative research is usually very small

and limited by the historical record, it is simply not possible to test an idea generated through in-depth study using “another sample” of cases. Very often, the cases included at the outset of an investigation comprise the entire universe of relevant cases.

The immediate objective of most comparative research is to explain the “how” of historically or culturally significant phenomena—for example, How do guerilla movements form? Theory plays an important orienting function by providing important leads and guiding concepts for empirical research, but existing theory is rarely well formulated enough to provide explicit hypotheses in comparative research. The primary scientific objective of comparative research is not theory testing, *per se*, but concept formation, elaboration, and refinement. In comparative research, the bulk of the research effort is often directed toward constituting the cases in the investigation and sharpening the concepts appropriate for the cases selected (Ragin & Becker, 1992).

THE DEPENDENT VARIABLE

One of the most fundamental notions in textbook presentations of social research is the idea of the *variable*—a trait or aspect that varies from one case to the next—and the associated idea of looking for patterns in how variables are related across cases. For example, do richer countries experience less political turmoil than poorer countries? If so, then social scientists might want to claim that variation in political turmoil across countries (the dependent or outcome variable) is explained in part by variation in country wealth (the independent or causal variable). Implicit in these notions about variables is the principle that the phenomena that social scientists wish to explain must *vary* across the cases they study; otherwise, there is nothing to explain. It follows that researchers who study outcomes that vary little from one case to the next have little hope of identifying the causes of those outcomes.

Comparative researchers, however, often intentionally select cases that do not differ from each other with respect to the outcome under investigation. For example, a researcher might attempt to explain how guerilla movements come about and study instances of their successful formation. From the viewpoint of textbook social research, however, this investigator has committed a grave mistake—selecting cases that vary only slightly, if at all, on the dependent variable. After

all, the cases are all instances of successfully formed guerilla movements. Without a dependent variable that varies, this study seems to lack even the possibility of analysis.

These misgivings about studying instances of “the same thing” are well reasoned. However, they are based on a misunderstanding of the logic of comparative research. Rather than using variation in one variable to explain variation in another, comparative researchers often look for causal conditions that are common across their cases—that is, across all instances of the outcome (see Robinson, 1951). For example, if all cases of successfully formed guerilla movements occur in countries with both high levels of agrarian inequality and heavy involvement in export agriculture, then these structural factors might be important to the explanation of guerilla movements. In fact, the very first step in the comparative study of successfully formed guerilla movements, after constituting the category and specifying the relevant cases, would be to identify causal conditions that are common across instances. Although using constants to account for constants (i.e., searching for causal commonalities shared by similar instances) is common in comparative research, it is foreign to techniques that focus on relationships among variables—on causal conditions and outcomes that vary substantially across cases.

Sometimes, it is possible for comparative researchers to identify relevant negative cases to compare with positive cases. For example, guerilla movements that failed early on might be used as negative cases in a study of successfully established guerilla movements. Negative cases are essential when the goal is to identify the *sufficient* conditions for an outcome (Ragin, 2000). For example, if a comparative researcher argues that a specific combination of causal conditions is sufficient for the successful establishment of a guerilla movement, it is important to demonstrate that these conditions were not met in the near misses (i.e., in relevant negative cases). Note, however, that it is possible to assess the sufficiency of a set of causal conditions only if the researcher is able to constitute the set of relevant negative cases. Often, it is very difficult to do so, either because the empirical category is too broad or vague (e.g., the category “absence of a successful guerilla movement”) or the evidence is impossible to locate or collect (e.g., data on aborted movements). When these obstacles can be overcome, it is important to understand that the

constitution and analysis of the positive cases is a necessary preliminary for the constitution and analysis of negative cases (see also Griffin et al., 1997). Thus, even when a comparative researcher plans to look at an outcome that varies across cases (i.e., present in some and absent in others), it is often necessary first to study cases that all have the outcome (i.e., the positive cases).

CAUSAL ANALYSIS

The main goal of most analyses of social data, according to textbook presentations of the logic of social research, is to assess the relative importance of independent variables as causes of variation in the dependent variable. For example, a researcher might want to know which has the greater impact on the longevity of democratic institutions, their design or their perceived legitimacy. In this view, causal variables compete with each other to explain variation in an outcome variable. A good contender in this competition is an independent variable that is strongly correlated with the dependent variable but has only weak correlations with its competing independent variables.

Comparative researchers, by contrast, usually look at causes in terms of combinations: How did relevant causes combine in each case to produce the outcome in question? Rather than viewing causes as competitors, comparative researchers view them as raw ingredients that combine to produce the qualitative outcomes they study. The effect of any particular causal condition depends on the presence and absence of other conditions, and several different conditions may satisfy a general causal requirement—that is, two or more different causes may be equivalent at a more abstract level. After constituting and selecting relevant instances of an outcome (e.g., successfully formed guerilla movements) and, if possible, defining relevant negative cases as well, the comparative investigator's task is to address the causal forces behind each instance, with special attention to similarities and differences across cases. Each case is examined in detail, using theoretical concepts, substantive knowledge, and interests as guides to answer "how" the outcome came about in each positive case and why it did not in the negative cases (assuming they can be confidently identified). A common finding in comparative research is that different combinations

of causes may produce the same outcome (Ragin, 1987; see also Mackie, 1974).

SUMMARY

The contrasts sketched in this entry present a bare outline of the distinctive features of comparative research, from the constitution of cases to the examination of the different combinations of conditions linked to an outcome across a set of cases. The key point is that comparative research does not conform well to textbook admonitions regarding the proper conduct of social research. Social scientists and their audiences are often interested in understanding social phenomena that are culturally or historically significant in some way. Of necessity, these phenomena are often rare and can be understood only through in-depth, case-oriented research.

—Charles C. Ragin

See also CASE STUDY

REFERENCES

- Eckstein, H. (1975). Case study and theory in political science. In F. I. Greenstein & N. W. Polsby (Eds.), *Handbook of political science: Vol. 7. Strategies of inquiry* (pp. 79–137). Reading, MA: Addison-Wesley.
- Griffin, L., Caplinger, C., Lively, K., Malcom, N. L., McDaniel, D., & Nelsen, C. (1997). Comparative-historical analysis and scientific inference: Disfranchisement in the U.S. South as a test case. *Historical Methods*, 30, 13–27.
- Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford, UK: Oxford University Press.
- Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.
- Ragin, C. C. (1994). *Constructing social research*. Thousand Oaks, CA: Pine Forge Press.
- Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago: University of Chicago Press.
- Ragin, C. C., & Becker, H. S. (1992). *What is a case? Exploring the foundations of social inquiry*. New York: Cambridge University Press.
- Robinson, W. S. (1951). The logical structure of analytic induction. *American Sociological Review*, 16, 812–818.
- von Wright, G. H. (1971). *Explanation and understanding*. Ithaca, NY: Cornell University Press.
- Walton, J. (1992). Making the theoretical case. In C. C. Ragin & H. S. Becker (Eds.), *What is a case? Exploring the foundations of social inquiry* (pp. 121–137). New York: Cambridge University Press.

COMPARATIVE RESEARCH

The major aim of comparative research is to identify similarities and differences between social entities.

Comparative research seeks to compare and contrast nations, cultures, societies, and institutions. Scholars differ on their use of the terminology: To some, comparative research is strictly limited to comparing two or more nations (also known as “cross-national research”), but other scholars prefer to widen the scope to include comparison of many different types of social and/or cultural entities. Yet other scholars use the term to encompass comparisons of subcultures or other social substrata either within or across nation-states or other cultural and social boundaries.

Although scholars are far from a consensus on a definition, the trend appears to be toward defining comparative research in the social sciences as research that compares systematically two or more societies, cultures, or nations. In actual implementation, comparisons of nations prevail as the dominant practice. The intent of comparative research is more universal: Comparative research aims to develop concepts and generalizations based on identified similarities and differences among the social entities being compared, especially in their characteristic ways of thinking and acting; in their characteristic attitudes, values, and ideologies; and in the intrinsic elements of their social structures. This then serves as a means of enhancing one’s understanding and awareness of other social entities.

HISTORICAL DEVELOPMENT

Sociology’s founding fathers were all comparative researchers. Karl Marx, Max Weber, Emile Durkheim, and Alexis de Tocqueville, to name just a few, all firmly committed themselves to the comparative method, whether they studied roles, institutions, societies, nations, cultures, groups, or organizations. Durkheim, for example, sought to find a general “law” that would explain national and occupational suicide rate variations. In the early 20th century, though, comparative research waned because its methodological resources were deemed insufficiently rigorous during a time when scholars were insisting that social research conform to greater and greater levels of methodological precision. However, since World War II, comparative research has once again assumed its pivotal position

in social science research, due in part to improved methodologies and methodological tools and in part to the international climate that emerged after World War II. That climate (the once polarized communist and noncommunist worlds, for example) and our ensuing internationalization have both contributed to the significant reemergence of comparative social science research.

APPLICATION

The purposes of comparative research are many, but one key task is to support and contribute to theory formation. Although theoretical frameworks drive the construction of comparative research endeavors, the results of such research often drive theory reformation. Another key task is to support policymaking, principally at the national level. Can Country B use the strategic policies of Country A in coping with a given social problem (drug abuse, for example)? Or are there unique differences between the two countries that render such a strategy impractical? Yet another purpose of comparative research is to ascertain whether the same dimension of a given concept (e.g., religious commitment) can be used as a common social indicator. Does a given concept generalize to all nations (or other social entities)?

Comparative research is all about perspective. Researchers in one nation need a means to adopt and understand the perspectives of their counterparts in other nations. As the world becomes ever more globalized, the need for such understanding should be clear: National policies need to consider and encompass the needs of the global partners in any way affected by policy implementation. Do policymakers have the necessary perspective to ensure such cooperation and compliance?

Comparative social science research is implemented under a variety of methodologies, ranging from strictly qualitative to rigorously quantitative. Case studies and similar kinds of analyses typify the descriptive qualitative approach, which can be conducted by researchers who are members of the social entities being studied or by researchers from outside the social entities being examined. The quantitative approaches to comparative research have assumed a more dominant position in recent years as the tools (especially computational, statistical, and mathematical) have become available to deal with the many methodological complexities of rigorous comparative research.

Empirical research surveys have become quite common, despite their often enormous complexity and expense. Complexities include the language barrier, which makes preparing comparable questionnaires difficult, and “cultural” barriers, which often make the preparation of comparable questionnaires virtually impossible: If you ask Americans, “Do you believe in God?” almost everyone says yes, whereas if you ask the same question in Japan, only about one third say yes, meaning that, beyond identifying these consistent response percentages for each country, asking the question in the two environments makes little comparative sense because no meaningful implication can be inferred. Similarly, asking about employment versus unemployment is irrelevant to a social group in which there is no such market-based distinction. In this instance, there is no “conceptual equivalence” between the two social groups in terms of the notion of “employment.”

Recently, advances in statistical methods have significantly improved the social science researcher’s ability to analyze primary data and/or reanalyze secondary data. Methods such as COHORT ANALYSIS have been developed that can ease the need for rigorous sample selection, for instance, and greatly facilitate the detection and explication of similarities and differences among group data. Of course, like all rigorous scientific research, comparative social science research is subject to the usual requirements for statistical validity and reliability, among others.

Examples of ongoing comparative research surveys include the Gallup Polls (since 1945), the General Social Survey (since 1972), the Eurobaromètre (since 1973), the European Community Household Study (since 1994), and the International Social Survey Program (ISSP), which, since 1984, has conducted general social attitude surveys among more than 37 nations. The Japanese National Character Study has been carried out in Japan about every 5 years since 1953. Each of these surveys is conducted periodically over time and usually in multiple countries.

Why are nations different on some characteristic parameters but the same on others? The same question can be asked of other sets of sociocultural groupings, both across regional boundaries and within national or ethnic boundaries. Because national boundaries are not always the same as ethnic boundaries, comparative research is not always amenable to simple national boundaries. This phenomenon has been especially

relevant to the emerging countries of Eastern Europe in recent years.

The ultimate intent of comparative research is to contribute to theory building or rebuilding by requiring yet more and more rigor on the part of theory, by insisting that research be replicable across different social groups, by facilitating the application of generalizations across different social groups, and by explaining unique patterns.

FUTURE DIRECTIONS

If, as globalization implies, there is eventually social and cultural convergence in the world, comparative research will once again wane as there will be few distinct social entities to compare. In the meantime, however, there remains much variation and variability in the world, and comparative research promises to further explain this to us and to further enhance our understanding and awareness of one another.

—Masamichi Sasaki

REFERENCES

- Hantrais, L., & Mangen, S. (Eds.). (1996). *Cross-national research methods in the social sciences*. London: Pinter.
- Inkeles, A., & Sasaki, M. (Eds.). (1996). *Comparing nations and cultures*. Englewood Cliffs, NJ: Prentice Hall.
- Przeworski, A., & Teune, H. (1970). *The logic of comparative social inquiry*. New York: John Wiley.
- Smelser, N. (1997). *Problems of sociology*. Berkeley: University of California Press.

COMPARISON GROUP

In an experiment testing the effects of a treatment, a comparison group refers to a group of units (e.g., persons, classrooms) that receive either no TREATMENT or an alternative treatment. The purpose of a comparison group is to serve as a source of COUNTERFACTUAL causal inference. A counterfactual is something that did not happen (it is counter to fact); in this case, the counterfactual to receiving treatment is not receiving treatment. According to the theory of counterfactual causal inference, the effect of a treatment is the difference between the outcome of those who received a treatment and the outcome that would have happened had those same people

simultaneously not received treatment. Unfortunately, it is not possible to observe that counterfactual, that is, to both give and not give the treatment to the same people simultaneously. Therefore, researchers often use a comparison group as a substitute for the true but unobservable counterfactual, and then they estimate the treatment effect as the difference between the outcomes in the treatment group and the outcomes in the comparison group. To do this well, researchers should make sure that the units in the comparison group are as similar as possible to the units in the treatment group. That similarity can be facilitated by RANDOMLY ASSIGNING units to either treatment or control, resulting in a randomized experiment, or it can be facilitated less well by various QUASI-EXPERIMENTAL methods such as MATCHING units on variables thought to be related to outcome (Shadish, Cook, & Campbell, 2002).

KINDS OF COMPARISON GROUPS

A treatment group can be compared to many different kinds of comparison groups, for example, (a) a no-treatment condition, sometimes called as a CONTROL GROUP, in which the participants receive no intervention to alter their condition; (b) a treatment-as-usual condition, in which the standard care or treatment is provided; (c) a PLACEBO condition, in which the intervention mimics treatment but does not contain the active ingredients of that treatment; (d) a wait-list condition, in which participants are promised the same intervention that the treatment group received after the experiment ends; or (e) an alternative treatment condition, in which participants receive an intervention other than the one of primary interest in the experiment. In the latter case, the alternative treatment is often the best treatment currently available for a problem—a “gold standard”—to see whether the experimental treatment can improve on the outcomes it produces. A further distinction is that in BETWEEN-GROUPS comparisons, the comparison group is a different set of participants than the treatment group, but in WITHIN-SUBJECTS comparisons, the same participants serve sequentially in both the comparison and treatment groups. Within-subjects comparisons require fewer units to test a treatment but can suffer from various treatment contamination effects such as practice effects, fatigue effects, and CARRYOVER of treatment effects from one condition to another (Shadish et al., 2002).

HISTORICAL DEVELOPMENT

Comparison groups may have been used to aid causal inferences as early as ancient eras. Rosenthal and Rosnow (1991) relate a story of a magistrate in ancient Egypt who sentenced criminals to be executed by poisonous snakes. When the magistrate investigated why none of the criminals died after being bitten, he learned that they had been given citron to eat. To determine whether citron had prevented their fatalities, he fed citron to only half of the prisoners and subjected all prisoners to snake bites. In this example, the participants who were not given the citron were the comparison group, all of whom died after the snake bit, whereas none of those in the treatment group who received citron died.

EXAMPLES

Antoni et al. (1991) used a no-treatment control group to examine the effects of a cognitive-behavioral stress management treatment on stress levels of homosexual men prior to notifying them of their HIV status. Those in the treatment group received training in methods to reduce stress. Their stress levels were compared to stress levels in a no-treatment control group that did not receive such training. The treatment effect was the difference in stress levels between treatment and comparison groups after HIV status notification.

Mohr et al. (2000) used a treatment-as-usual comparison group to examine the effects of a cognitive-behavioral treatment to reduce depression in patients with multiple sclerosis (MS). The treatment group was given a new experimental treatment for depression, whereas the comparison group was given the usual treatment provided to MS patients suffering from depression. The resulting treatment effect is the difference in depression scores between those receiving standard treatments for depression compared to those receiving the new treatment for depression.

Kivlahan, Marlatt, Fromme, Coppel, and Williams (1990) used a wait-list control group to examine the effects of a skills program to reduce alcohol consumption among college students. The treatment was withheld from the comparison group until the end of the experiment, after which the comparison group was offered the opportunity to receive the same treatment given to the treatment group.

—William R. Shadish and M. H. Clark

See also EXPERIMENT, TREATMENT

REFERENCES

- Antoni, M. H., Baggett, L., Ironson, G., LaPerriere, A., August, S., Kilmas, N., et al. (1991). Cognitive-behavioral stress management intervention buffers distress responses and immunologic changes following notification of HIV-1 seropositivity. *Journal of Consulting and Clinical Psychology*, 59(6), 906–915.
- Kivlahan, D. R., Marlatt, G. A., Fromme, K., Coppel, D. B., & Williams, E. (1990). Secondary prevention with college drinkers: Evaluation of an alcohol skills training program. *Journal of Consulting and Clinical Psychology*, 58(6), 805–810.
- Mohr, D. C., Likosky, W., Bertagnolli, A., Goodkin, D. E., Van der Wende, J., Dwyer, P., et al. (2000). Telephone administered cognitive-behavioral therapy for the treatment of depressive symptoms in multiple sclerosis. *Journal of Consulting and Clinical Psychology*, 68(2), 356–361.
- Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

COMPETING RISKS

A set of mutually exclusive, discrete outcomes or events is called *competing risks*. This term arises primarily in the literatures on SURVIVAL ANALYSIS and EVENT HISTORY ANALYSIS.

As long ago as the 18th century, scholars interested in survival and its opposite (e.g., death, failure) recognized that deaths and failures occur for a variety of reasons. For example, a person may die from a contagious disease, an organic illness (e.g., cancer), an accident, suicide, or intentional violent action (e.g., murder, war). These alternative causes of death illustrate competing risks of dying. Frequently, scholars are interested not only in mortality and failure for all reasons combined but also in the diverse reasons for death or failure and in the associated cause-specific rates of mortality or failure. That is, they want to differentiate among competing risks of some event, such as death or failure.

Although the language and application of the idea of competing risks were originally developed in the fields of health, medicine, demography, engineering, and related fields concerned with death or failure, they apply to a broad spectrum of social scientific research topics. For example, researchers studying labor market

processes may wish to analyze not only exits from jobs overall but also whether people leave jobs because they have been promoted or demoted within an organization, moved laterally, quit, been fired, retired, or died. Scholars studying organizational success may want to examine not only all forms of organizational “death” but also whether organizations cease all activities, are acquired by another firm, or merge with another firm in something resembling an equal partnership. Social scientists studying international conflicts and wars may desire to examine not only the end of conflicts overall but also whether conflict ends through a negotiated peace process, by conquest, or with a stalemate. In each instance, scholars would like to distinguish among competing risks.

A recurring type of question concerns the consequences of altering the likelihood of one or more competing risks. A typical question is as follows: What would be the consequences of suppressing or completely eliminating a specific risk (e.g., a particular cause of death or failure)? To illustrate, the introduction of public sanitation and widespread immunization against contagious diseases greatly reduced the likelihood of catching contagious diseases. As a result, the risks and rates of dying from many of the most lethal contagious diseases were greatly diminished. Reduction in the rate of dying from any competing risk tends to extend life, causing an increase in the average length of life. More generally, lowering the HAZARD RATE of one competing risk prolongs the average time to the occurrence of an event for all reasons combined.

Competing risks of events are often highly age and time dependent. Consequently, the implications of a certain reduction in the hazard rate for one competing risk rather than for another can differ substantially, depending on the pattern of age and time dependence in each cause-specific hazard rate. For example, heart disease and cancer strike older people much more than children and youths, whereas contagious diseases and accidents befall infants and children more than adults. As a result, a certain reduction in the risk of dying from a contagious disease tends to lengthen life much more than a comparable reduction in the risk of dying from heart disease. Thus, improved sanitation and immunization greatly increased average life expectancy (and the length of life as a working-age adult) by causing the overall risk of dying at young ages to fall precipitously. In contrast, reducing the risk of dying from heart disease especially extends the remaining life of

the elderly and particularly increases the length of life spent in retirement.

An indirect consequence of a declining rate of death from contagious diseases that especially strike children was an increase in the relative likelihood of dying from heart disease, cancer, and other diseases to which elderly people are particularly susceptible. Improved sanitation and immunization did not eliminate other competing risks of dying; indeed, they may have not directly affected the rates of dying from these other causes at all. In general, direct suppression or elimination of one competing risk tends to increase the *relative* chances of death or failure due to competing risks other than the one that was suppressed or eliminated. Contrarily, amplification of one competing risk (e.g., murder, war) tends to decrease the *relative* chances of death or failure due to other competing risks (e.g., disease).

There are two main types of models for analyzing competing risks and the effects of measured covariates on the occurrence and timing of one competing risk rather than another. The two types are differentiated on the basis of whether risks may occur only at discrete points in time or at any continuous moment in time.

In the first instance, the most common statistical models are a MULTINOMIAL LOGISTIC REGRESSION MODEL or a conditional logistic regression model (cf. Theil, 1969, & McFadden, 1973, respectively, as cited in Maddala, 1983). In the second instance, event history analysis is the standard way of analyzing empirical data on competing risks. The specific models used in event history analysis include various types of proportional or nonproportional models of hazard and transitions rates. (For an overview of the main alternative models used in event history analysis, see Tuma & Hannan, 1984.)

A key issue in both types of modeling approaches is whether the competing risks are statistically independent. To give one example, a person who has an especially serious illness such as AIDS often has a heightened risk of dying from an infectious disease. If this is true, the two causes of death (e.g., AIDS and an infectious disease) are not statistically independent. In contrast, deaths from accidents, murder, and war may plausibly be considered to be independent (or nearly independent) of the risks of dying from a chronic illness such as AIDS.

When using discrete time multinomial logistic regression models to analyze the cause-specific occurrence of various outcomes, researchers may test the

independence of irrelevant alternatives (Debreu, 1960, as cited in Maddala, 1983), that is, whether distinct competing risks are statistically independent. If the competing risks are statistically independent, then estimates for a multinomial logistic regression model will be identical (within the limits of sampling variability) to the results for a series of simpler logistic regression models that exclude one or more of the competing risks. For a discussion of these tests, see Hausman and McFadden (1984).

In analyses of event histories, researchers may estimate the instantaneous transition rate (i.e., the cause-specific hazard rate) associated with each competing risk. The results of such analyses may then be used to infer the consequences of eliminating a particular competing risk, under the assumption that it is statistically independent of all other competing risks. Similarly, the same results may be used to infer the consequences of raising or lowering the transition rate associated with a certain risk (e.g., by multiplying the estimated transition rate by some hypothetical fraction greater than or less than 1). Such inferences are legitimate under the assumption that the competing risks are statistically independent (David & Moeschberg, 1978). Tsiatis (1975) proved that, even with full and complete event history data on the occurrence and timing of competing risks, it is impossible to establish conclusively that competing risks are truly statistically independent.

—Nancy Brandon Tuma

REFERENCES

- David, H. A., & Moeschberg, M. L. (1978). *The theory of competing risks* (Griffin's Statistical Monograph #39). New York: Macmillan.
- Hausman, J., & McFadden, D. (1984). Specification tests for the multinomial logit model. *Econometrica*, 52, 1219–1240.
- Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.
- Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72, 20–22.
- Tuma, N. B., & Hannan, M. T. (1984). *Social dynamics: Models and methods*. Orlando, FL: Academic Press.

COMPLEX DATA SETS

There are two distinct senses in which data are complex: statistical and structural. We can explain these

by contrasting them with a notional simple data set. This would consist of a set of entities corresponding to either an equal probability random sample of a population, in which each entity is independently selected, or a complete population. The data set contains the same complete set of measurements about each entity. For example, it might consist of a random sample of individuals who each respond to a set of questions about their political attitudes. Data sets are described as complex when they depart from these assumptions.

Statistical complexity arises when some of the standard assumptions about sampling do not hold, and standard statistical analyses may give misleading results without adjustment. Skinner, Holt, and Smith (1989) give an overview of this sort of statistical complexity. One example of a departure occurs if, in a survey, respondents are clustered in a limited set of areas, as is often done to reduce fieldwork costs. Observations are not then independent. This will mean that standard errors used to compute significance tests will need to be adjusted compared to the standard formulas used. For example, the British Social Attitudes Survey, which might at first sight appear to be the type of simple data set envisaged in the first paragraph, has a clustered design. Other forms of lack of independence also occur in spatial and time-series data sets. Another departure occurs if the sample is not drawn with equal probability. Estimates will then be biased and need to be adjusted, for example, by weighting. This is quite typical in surveys of employers, in which large employers tend to be selected with higher probability. Similarly, if there are variations in the probabilities of response by sampled individuals, either to the whole interview or to individual questions, then there may be biases that require adjustment.

Structural complexity arises when we have information about different sorts of entities in a single data set. In particular, many data sets have a multilevel structure, reflecting some important aspect of social reality. Examples include not only hierarchical structures, such as individuals located in households, workers in firms, and children in schools, but also multiple observations in longitudinal data of the same individual at different points in time. The collection of information about all individuals in households is quite typical in household budget or labor force surveys. There are other less clearly hierarchical structures, for example, in the study of social networks. Data sets may also be seen as complex when there are different amounts of information about each primary entity—for example,

different numbers of observations in a panel study, different numbers of children born to each woman in a fertility history, or different numbers of jobs held in a work history. For instance, a survey such as the British Household Panel Survey will have a different number of waves for different individuals both because of attrition and because children reach the eligible age for interview.

There are a number of consequences of this structural complexity. First, it means that analysts have to make choices about what their unit of analysis should be. It is possible with these structures to treat higher level units as the unit of analysis and attach summary information about lower level units (e.g., household-level analysis, using information about the income of individuals in the household or analyzing women by number of children born). Alternatively, it is possible to use a lower level for analysis and attach summary information about the higher level (e.g., analysis of each separate birth). Second, given these choices, the data may require substantial restructuring to create the rectangular data sets normally used by statistical software. Finally, analysts must take account of statistical consequences of the structural complexity. This might be simply to adjust for nonindependence of observations, for example, but more positively, it may involve taking advantage of the structure by using multilevel modeling techniques (see Goldstein, 1995).

—Nicholas H. Buck

REFERENCES

- Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.
- Skinner, C., Holt, D., & Smith, T. M. F. (1989). *Analysis of complex surveys*. Chichester, UK: Wiley.

COMPONENTIAL ANALYSIS

Componential analysis (CA) assumes that the meaning of any given word is represented best by a unique bundle of meaningful features. The analytical method of CA is to compare the meanings of words from the same area of meaning to discover the minimum number of features necessary to distinguish the differences in their meanings. For example, in the area of kinship, a person's *mother* and *father*; *sisters* and *brothers*, and *daughters* and *sons* are distinguished

by gender (“female” vs. “male”) and by generation (“one generation earlier” for *mother* and *father*; “same generation” for *sister* and *brother*; and “one generation later” for *daughter* and *son*).

The insights gained through CA are regarded as revealing how the people of different societies uniquely structure their knowledge of the world and interpret life experience. CA is therefore regarded as one of the main tools for addressing the nature of cultural and linguistic relativism.

Proponents of CA purport that it provides an insightful, elegant, and rigorous method for determining lexical meanings. It is insightful in that it reveals unexpected bases for meaningful relations, elegant in that it exhibits a simplicity and regularity not available through other methods, and rigorous in that it provides precise definitions. Most dictionary definitions are composed of such features (e.g., *trudge* “to walk in a laborious and weary manner”).

Proponents of CA also hope that it will provide a set of universal features that will allow for the analysis of any meaning in any language, thereby overcoming imprecision when translating between languages. The translator may compensate for the difference in the meanings of words from different languages by adding phrases to account for missing meaningful features. For example, when another language has several words for specific ways to *carry* something, one may express those differences by adding to *carry* phrases, such as “in the hands,” “on the hip,” “under the arm,” “over the shoulder,” and so forth.

From the mid-1950s through the 1970s, CA played a major role in the development of the discipline of COGNITIVE ANTHROPOLOGY (D’Andrade, 1995). It was subsequently rejected by the academic community, however, for a number of reasons. First, it allows for alternative analyses that cannot be evaluated in terms of how well each reflects a cultural insider’s knowledge. Second, it cannot account for the meanings of a wide range of domains. For example, words that reflect a part-whole relationship, such as the *head* as a part of the body, may only be understood in relation to their position relative to the other parts of the whole rather than by a list of features (Fillmore, 1975). Third, categories that lack clearly defined boundaries defy analysis by CA. Examples include the stages of life (*infant*, *child*, *adolescent*, *adult*, and *elder*) and size (*sand*, *gravel*, *pebbles*, *stones*, *rocks*, and *boulders*). Fourth, the meanings of many words can only be understood in terms of the broader cultural context. For

example, a *fiddle* can be discriminated from a *violin* because it may be grouped with a guitar, ukulele, and a banjo and played to accompany folk dancing at a country hoedown by men dressed in bib overalls and plaid shirts. A violin may be grouped with another violin, a viola, and a cello as part of a string quartet and played at a chamber music concert by players dressed in tuxedos and gowns.

The enduring value of CA is that it highlights the fact that people discriminate things on the basis of whether they are the same or different.

—Kenneth A. McElhanon and Thomas N. Headland

See also COGNITIVE ANTHROPOLOGY, EMIC/ETIC DISTINCTION

REFERENCES

- D’Andrade, R. (1995). *The development of cognitive anthropology*. Cambridge, UK: Cambridge University Press.
- Fillmore, C. (1975). An alternative to checklist theories of meaning. In C. Cogen, H. Thompson, G. Thurgood, K. Whistler, & J. Wright (Eds.), *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society* (pp. 123–131). Berkeley, CA: Berkeley Linguistics Society.
- Nida, E. A. (1975). *Componential analysis of meaning: An introduction to semantic structures*. The Hague, The Netherlands: Mouton.

COMPUTER SIMULATION

A computer simulation uses a set of equations, ALGORITHMS, or electronic components to reproduce the essential components of a more complex system. Its purpose is to allow the complex system to be studied under conditions that would be difficult, expensive, or impossible to manipulate in the real world. Computer simulations also allow for mathematical equations to be studied numerically, even when they cannot be solved using algebraic techniques.

Most computer simulations are implemented on digital computers and specified using equations, if-then rules, and RANDOM VARIABLES. A variety of specialized computer languages and packages have been developed for the specification of simulations, although simulations can also be written in general-purpose programming languages such as FORTRAN or C++. The behavior of the simulation will vary depending on the PARAMETERS used in the equations and rules, the

values used to initialize the simulation, and the values of the random variables during a specific run of the simulation.

Most computer simulations fall into three categories: systems dynamics simulations, finite element simulations, and agent-based simulations.

Systems dynamics simulations, pioneered by Forrester (1961), are driven by a set of difference equations that relate the variables in the system through a set of feedback loops and random disturbances; Roberts, Andersen, Deal, Grant, and Shaffer (1983) provide an extended introduction. For example, "global dynamics" simulations study highly aggregated variables such as the total level of population, resource consumption, pollution, and economic production. Pollution would increase as production and consumption increase, which might, in turn, reduce the level of population.

Finite element simulations, in contrast, compute the behavior of aggregate variables by looking at the effects of microlevel processes occurring between adjacent elements in the simulation. This approach is typically used to model physical systems, such as climate, in which the microlevel equations are known but the systems cannot be solved algebraically due to the presence of random variables and NONLINEAR equations. For example, climate change models contain equations governing the effects of greenhouse gases such as carbon dioxide on heat retention in the atmosphere, the effect of atmospheric heat on adjacent ocean temperature, the effects of ocean temperature on cloud formation in the atmosphere above the ocean, and so forth. Most finite element simulations can be efficiently implemented on massively parallel computers that contain hundreds or thousands of individual processors, and the most common application of these "supercomputers" is finite element simulation.

Finally, agent-based simulations model social systems by looking at the aggregated effects of individual agents whose behaviors are specified by rules (Axelrod, 1997). For example, in a simulation of traffic flow in a city, the "agents" would be individual cars, which would alter their speed, choice of route, and random behaviors such as the likelihood of being in an accident based on the actions of other agents. See the SIMULATION entry for further details on this approach.

Although most simulations of social systems are done with digital computers, analog computer simulations are occasionally used in MODELING in the social sciences, particularly for aspects of the nervous system and tasks such as pattern recognition. An analog

computer is constructed from electronic components such as resistors, capacitors, and transistors. Like a digital simulation, an analog computer is a simplification of a more complex system, but its characteristics can be more easily and reliably manipulated than can be done with the full system.

—Philip A. Schrodtt

REFERENCES

- Axelrod, R. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton, NJ: Princeton University Press.
- Forrester, J. W. (1961). *Industrial dynamics*. Cambridge: MIT Press.
- Roberts, N., Andersen, D. F., Deal, R. M., Grant, M. S., & Shaffer, W. A. (1983). *Introduction to computer simulation: A system dynamics modeling approach*. Reading, MA: Addison-Wesley.

COMPUTER-ASSISTED DATA COLLECTION

This term can be used in relation to several different kinds of context: (a) face-to-face and telephone interviews, in which the interviewer follows a series of questions from an INTERVIEW SCHEDULE that is on a computer and enters respondents' replies as they go along; (b) SELF-ADMINISTERED QUESTIONNAIRES, in which respondents read the questions on a computer screen and enter their replies as they go along; (c) INTERNET SURVEYS; and (d) online FOCUS GROUPS and personal interviews.

—Alan Bryman

COMPUTER-ASSISTED PERSONAL INTERVIEWING

Computer-assisted personal interviewing (CAPI) and its cousin, computer-assisted telephone interviewing (CATI), dominate survey data collection (see COMPUTER-ASSISTED DATA COLLECTION). There are a large number of software packages available and more coming online. In this entry, we will not attempt to evaluate the alternatives but instead discuss the major technical problems any CAPI system needs to solve

and provide the outline of a solution. The advantages of computer-assisted interviewing are so substantial—such as efficient question filters, range checks, and skip pattern checks—that the only candidates for *not* using computer-assisted techniques are very short, simple questionnaires or surveys that involve few respondents or can use low-wage labor to process data and cannot afford the hardware.

Mounting a major SURVEY requires the coordination of many tasks and details that touch data collection. Devising a CAPI software system to handle a large, complex survey is difficult. Unless the CAPI system integrates well with the other aspects of the survey process, the benefits of computerized data collection can be not only offset but also outweighed by the difficulties created by a poorly integrated effort. CAPI initially promised more accurate data collection through built-in checks that prevent or correct errors and inconsistencies. It also promised faster data delivery times from the end of the field effort to the release of the data. CAPI has met these promises. On the other hand, organizations that paid for surveys hoped that CAPI would reduce their costs. In general, this has not happened. We believe that problems with cost reflect a failure to integrate CAPI software with the rest of the survey process.

In mounting a major CAPI effort, there are six closely linked tasks. First, we must design the interview to meet the needs of the project. Second, we must prepare the software that puts the question and, where appropriate, the set of allowable answers on the computer screen. Third, we prepare any input data that drives the interview to the software. The input data can range from the rudimentary—with little or no data about the respondent being fed into the survey—to large and complex data for longitudinal surveys that contain hundreds or even thousands of data points from previous rounds of the study (which can be referenced during the interview). For example, when collecting EVENT HISTORIES, the names and demographic characteristics of household members, names of employers, and dates and types of transitions figure prominently in the input file to a longitudinal survey. Input data can also include sound files for applications in which the computer “reads” the question text to the respondent, a technique used for sensitive questions about sexual practices or substance use when directly reporting to the interviewer might bias the response. The fourth task is to train interviewers to use the software and send them out to collect or “capture” the data. The

fifth task is to move the data from the interviewer’s portable PC to a central data repository; the sixth task is to prepare data files for analysis from the data in that central repository.

Although the balance of this entry will focus on software, any CAPI effort needs hardware to run on. Currently, there are three categories: laptops, pen-based systems, and small handheld personal digital assistants (PDAs). For large, complex surveys, laptops are currently the best choice for variety, robustness, stability, and product maturity. However, pen-based systems based on Windows/Intel technology are making another play for market share. Earlier pen-based systems did not fare well. The pen-based Windows/Intel machines are distinctly better than earlier generations, but for a notoriously cost-conscious industry, price, reliability, and physical robustness are the keys to accepting the new generation of pen-based PCs for survey use. For short surveys done standing up, the new vintage of PDAs that have more processing power and greater storage space could be strong contenders, but their relatively immature operating systems and smaller collection of software utilities are detrimental to wider use. However, for some applications in which small size, light weight, and modest cost are prerequisites, PDAs are already an excellent choice.

In the effort to write software to handle data collection, systems designers tended to focus on the data capture step. Often, they did not pay sufficient attention to how system design might affect interviewer training, as well as how the data capture software integrated with the other five tasks. With the advent of CAPI, the entire back office operation for survey work has had to be redesigned. Without an overarching conceptual framework for how the major tasks for CAPI data collection should be integrated, the costs of the redesigned survey process have often taken an unfortunate direction.

CAPI/CATI technologies have two major branches. The more common approach uses a customized programming language to present questions, capture answers, and store data. This approach requires extensive support from programmers throughout the survey process, and for every new survey, the interrelationships among these tasks must be specified anew. We will refer to this as the *programming approach*. The other approach to CAPI/CATI software uses relational database methods as a unifying mechanism. With this approach, the instrument is represented as a series

of data records, each of which has a large number of attributes, including everything from the question text and skip pattern to documentation notes and instructions to the interviewer. We will refer to this as the *database approach*. (The reader can profitably read the entry on DATA MANAGEMENT, which explains relational databases, as it is germane to this section as well.) The thesis here is that the database approach provides an overarching conceptual approach to integrating all the steps in conducting a CAPI interview and allows for significant operational economies over the alternative programming approach.

In the design stage, survey specialists enter data into the various fields of a screen that populates the rows of relational database tables representing the questionnaire. The data entered include question text, edit restrictions on allowable responses, the list of acceptable answers, and skip instructions. This approach to “authoring” the CAPI questionnaire requires minimal programming skill, and junior survey specialists can enter all but the most complex passages. A set of pre-designed “queries” to the database generates diagnostic checks as the data are entered, with more sophisticated checks and diagnostics run in batch mode.

Each question in the instrument is represented by joining rows from the database tables. The interviewing software performs these joins, displays the question, stores the answer, and branches as per the branching instructions. The program that reads and executes the questionnaire data is not rewritten for each survey but remains stable and is reused for different surveys. (The same records can also be used on a PDA, Web server, or local server to enable a variety of platforms for data collection.) This simplifies training for both questionnaire authors and interviewers because all interactions with the system reuse a few standard screen displays.

The survey data are transmitted to the central office and loaded into a master relational database already containing the tables that define the questionnaire. At this point, the questionnaire data become the “meta-data,” or data that describe the survey data, greatly facilitating the generation of documentation reports that are run as queries to the relational database. Pre-designed queries generate either a printable or HTML questionnaire or a codebook. Additional information relevant to each question can be stored in the database and used to augment survey documentation.

We can extract input data files for the next round of a longitudinal survey from the relational database.

One can distribute the public use data over the Web to users by employing the relational database techniques used for commercial transactions; instead of putting books in their shopping cart, researchers can search the database for the variables they need, mark them for extraction, and have a remote server e-mail them their data file.

This approach is holistic and uses commercial database software to integrate all phases of a CAPI survey. A well-designed system will handle a broad range of surveys, allowing both central office and interviewing staff to reuse standard tools. This approach offers significant opportunities to control costs and keep operations coordinated and efficient (see DATA MANAGEMENT).

—Randall J. Olsen

REFERENCES

- Costigan, P., & Thomson, K. (1992). Issues in the design of CAPI questionnaires for complex surveys. In A. Westlake, R. Banks, C. Payne, & T. Orchard (Eds.), *Survey and statistical computing* (pp. 47–156). London: North Holland.
- Couper, M. P., Baker, R. P., Bethlehem, J., Clark, C. Z. F., Martin, J., Nichols, W. L., et al. (Eds.). (1998). *Computer assisted survey information collection*. New York: John Wiley.
- Forster, E., & McCleery, A. (1999). Computer assisted personal interviewing: A method of capturing sensitive information. *IASSIST Quarterly*, 23(2), 26–38.
- Saris, W. E. (1991). *Computer-assisted interviewing*. Newbury Park, CA: Sage.

CONCEPT. See CONCEPTUALIZATION, OPERATIONALIZATION, AND MEASUREMENT

CONCEPTUALIZATION, OPERATIONALIZATION, AND MEASUREMENT

Research begins with a “problem” or topic. Thinking about the problem results in identifying concepts that capture the phenomenon being studied. Concepts, or CONSTRUCTS, are ideas that represent the phenomenon. Conceptualization is the process whereby these concepts are given theoretical meaning. The process typically involves defining the concepts abstractly in theoretical terms.

Describing social phenomena and testing hypotheses require that concept(s) be operationalized. Operationalization moves the researcher from the abstract level to the empirical level, where variables rather than concepts are the focus. It refers to the operations or procedures needed to measure the concept(s). Measurement is the process by which numerals (or some other labels) are attached to levels or characteristics of the variables. The actual research then involves empirically studying the variables to make statements (descriptive, relational, or causal) about the concepts.

CONCEPTUALIZATION

Although research may begin with only a few and sometimes only loosely defined concepts, the early stages usually involve defining concepts and specifying how these concepts are related. In exploratory research, a goal is often to better define a concept or identify additional important concepts and possible relationships.

Because many concepts in social science are represented by words used in everyday conversation, it is essential that the concepts be defined. For example, the concepts *norms*, *inequality*, *poverty*, *justice*, and *legitimacy* are a part of our everyday experiences and thus carry various meanings for different people. The definitions provided by the researchers are usually referred to as nominal definitions (i.e., concepts defined using other words). No claim is made that these definitions represent what these concepts “really” are. They are definitions whose purpose it is to communicate to others what the concept means when the word is used. A goal in social science is to standardize definitions of the key concepts. Arguably, the most fruitful research programs in social science—those that produce the most knowledge—are those in which the key concepts are agreed on and defined the same way by all.

Concepts vary in their degree of abstractness. As an example, human capital is a very abstract concept, whereas education (often used as an indicator of human capital) is less abstract. Education can vary in quality and quantity, however, and thus is more abstract than years of formal schooling. Social science theories that are more abstract are usually viewed as being the most useful for advancing knowledge. However, as concepts become more abstract, reaching agreement on appropriate measurement strategies becomes more difficult.

OPERATIONALIZATION

Operationalization of concepts involves moving from the abstract to the empirical level. Social science researchers do not use this term as much as in the past, primarily because of the negative connotation associated with its use in certain contexts. One such use has been in the study of human intelligence. Because consensus on the meaning of this concept has been hard to come by, some researchers simply argued that intelligence is what intelligence tests measure. Thus, the concept is defined by the operations used to measure it. This so-called “raw empiricism” has drawn considerable criticism, and as a consequence, few researchers define their concepts by how they are operationalized. Instead, nominal definitions are used as described above, and measurement of the concepts is viewed as a distinct and different activity. Researchers realize that measures do not perfectly capture concepts, although, as described below, the goal is to obtain measures that validly and reliably capture the concepts.

MEASUREMENT

Measurement refers to the process of assigning numerals (or other labels) to the levels or characteristics of variables. In moving from the abstract level of concepts to the empirical level of variables, the researcher must think of the concept as having characteristics that can be empirically observed or assessed. A number of concerns must be addressed. The first of these is LEVEL OF MEASUREMENT.

At the most basic level, the researcher must decide whether the underlying features of the concept allow for ordering cases (ordinal level) or allow only for categorizing cases (nominal level). Another distinction concerns whether the features of the concept are discrete or continuous, with fine gradations. Relying on these distinctions, most researchers are faced with variables that are nonordered and discrete (e.g., marital status and religious preference), ordered and discrete (e.g., social class), or ordered and continuous (e.g., income). Developments in statistics over the past several decades have dramatically increased the analysis tools associated with ordered discrete variables, but the majority of the research in social science still adheres to the dichotomy of categorical versus continuous variables, with the more “liberal” researchers assuming that ordinal-level data can be analyzed as if they were continuous. With the

advances in statistical techniques, however, these decisions no longer need to be made; ordinal variables can be treated as ordinal variables in sophisticated multivariate research (see Long, 1997).

The first step in measurement, then, is to determine the level of measurement that is inherent to the concept. As a general rule, measurement should always represent the highest level of measurement possible for a concept. The reason for this is simple. A continuous variable that is measured like a category cannot be easily, if ever, converted to a continuous scale. A continuous scale, however, can always be transformed into an ordinal- or nominal-level measure. In addition, ordinal and continuous scales allow for the use of more powerful statistical tests, that is, tests that have a higher likelihood of rejecting the null hypothesis when it should be rejected.

Any measurement usually involves some type of measurement ERROR. One type of error, *measurement invalidity*, refers to the degree to which the measure incorrectly captures the concept (DeVellis, 1991). Invalidity is referred to as systematic error or bias. VALIDITY is usually thought of in terms of the concept being a target, with the measure being the arrow used to hit the target. The degree to which the measure hits the center of the target (the concept) is the measure's validity. Researchers want measures with high validity. The reason for this is simple, as shown by an example from a study of social mobility. If the researcher's goal is to determine the degree of social mobility in a society, and the measure consistently underestimates mobility, then the results with this measure are biased and misleading (i.e., they are invalid).

The other major type of measurement error is *unreliability*. Whereas invalidity refers to accuracy of the measure, unreliability refers to inconsistency in what the measure produces under repeated uses. Measures may, on average, give the correct score for a case on a variable, but if different scores are produced for that case when the measure is used again, then the measure is said to be unreliable. Fortunately, unreliability can be corrected for statistically; invalidity is not as readily corrected, however, and it is not as easily assessed as is RELIABILITY.

Validity is arguably the more important characteristic of a measure. A reliable measure that does not capture the concept is of little value, and results based on its use would be misleading. Careful conceptualization is critical in increasing measurement validity. If the concept is fuzzy and poorly defined, measurement

validity is likely to suffer because the researcher is uncertain as to what should be measured. There are several ways of assessing measurement validity.

FACE VALIDITY, whether the measure "on the surface" captures the concept, is not standardized or quantifiable but is, nevertheless, a valuable first assessment that should always be made in research. For example, measuring a family's household income with the income of only the male head of household may give consistent results (be reliable), but its validity must be questioned, especially in a society with so many dual-earner households.

Content validity is similar to face validity but uses stricter standards. For a measure to have content validity, it must capture all dimensions or features of the concept as it is defined. For example, a general job satisfaction measure should include pay satisfaction, job security satisfaction, satisfaction with promotion opportunities, and so on. As with face validity, however, content validity is seldom quantified.

CRITERION-RELATED VALIDITY is the degree to which a measure correlates with some other measure accepted as an accurate indicator of the concept. This can take two forms. *Predictive validity* uses a future criterion. For example, voting preference (measured prior to the election) is correlated with actual voting behavior. A high correlation indicates that voting preference is a valid measure of voting behavior. *Concurrent validity* is assessed by obtaining a correlation between the measure and another measure that has been shown to be a valid indicator of the concept. For example, a researcher in the sociology of work area wishes to measure employee participation by asking employees how much they participated in decision making. A more time-consuming strategy—videotaping work group meetings and recording contributions of each employee—would likely be viewed as a valid measure of participation. The correlation between the two would indicate the concurrent validity of the perceptual measure.

CONSTRUCT VALIDITY of a measure refers to one of two validity assessment strategies. First, it can refer to whether the variable, when assessed with this measure, behaves as it should. For example, if the theory (and/or past research) says it should be related positively to another variable *Y*, then that relationship should be found when the measure is used. The second use of construct validity refers to the degree to which multiple indicators of the concept are related to the underlying construct and not to some other

construct. Often, this distinction is referred to with the terms *convergent validity* and DISCRIMINANT VALIDITY. FACTOR ANALYSIS can be used to assess this. For example, if a researcher has five indicators of cultural capital and four indicators of social capital, a factor analysis should produce two lowly correlated factors, one for each set of indicators.

It is important to stress that this second strategy uses factor analysis for confirmatory purposes, not for data dredging or concept hunting, as it is used sometimes in exploratory research. If the researchers have been careful in conceptualization, then the factor analysis serves to validate their measures. Factor analysis in this instance is not used to “discover” unanticipated concepts.

As stated above, reliability refers to the consistency of the results when the measure is used repeatedly. It is viewed as random measurement error, whereas validity is thought of as measurement bias. There are clear statistical consequences of unreliable measures, with larger variances and attenuated correlations at the top of the list. Reliability is usually assessed in one of two ways. TEST-RETEST RELIABILITY is assessed with the correlation of the measure with itself at two points in time. The difficulty associated with this type of assessment is that unreliability of the measure is confounded with actual change in the variable being measured. For this reason, a second and more often used strategy is to obtain multiple indicators of the concept. Here, internal consistency measures such as the CRONBACH'S ALPHA can be used. In both instances, a normed measure of association (correlation) allows for application of certain rules of thumb. For example, any measures with reliabilities under .60 should be regarded with considerable suspicion.

ADDITIONAL ISSUES

Measurement validity should not be confused with INTERNAL VALIDITY and EXTERNAL VALIDITY, which are research design issues. If researchers claim the results support their hypothesis that *X* causes *Y*, this claim is said to have internal validity if alternative explanations for the results have been ruled out. External validity refers to whether the study results can be generalized beyond the setting, sample, or time frame for the study. Both are very important concerns in building social science knowledge, but they are not the same as measurement validity.

Another measurement issue concerns whether single or multiple indicators should be used as measures. Some concepts, almost by definition, can be measured with single indicators. Some examples are income and education. Other more abstract concepts (e.g., political conservatism and marital satisfaction) are often measured with more than one indicator. The reasons for wanting multiple indicators are fairly straightforward. First, when a concept is more abstract, finding one measure that captures it is more difficult. Second, multiple-indicator measures are usually more reliable than single-indicator measures. Third, multiple-indicator measures, when used in STRUCTURAL EQUATION MODELING (SEM) (Bollen, 1989), allow for correcting for unreliability of the measures, produce additional information that allows for making useful model tests, and allow for estimating more complicated models such as reciprocal effects models. Although SEM achieves its strengths by treating multiple indicators separately, measures based on multiple indicators are often transformed into scales or indexes, which are the simple sums or weighted sums of the indicators.

A third measurement issue concerns the source of valid and reliable measures. As has been described, there exist a number of strategies for assessing how good a measure is (i.e., how valid and reliable it is). This assessment happens after the measurement has occurred, however. If the measure is not valid or reliable and other better measures cannot be easily obtained, discontinuation of the research must be seriously considered. For this reason, considerable attention must be given to identifying valid and reliable measures at the onset of the study. If the theory in a specialty area is well established and there already exists a strong research tradition, then valid and reliable measures likely already exist. For example, in the study of organizations, key concepts such as formalization and administrative intensity are well established, as are their measures. In such instances, a careful reading of this literature identifies the measures that should be used. It is always tempting for those new to a specialty area to want to “find a better measure,” but unless previously used measures have low validity and reliability, this temptation should be avoided. Accumulation of knowledge is much more likely when measures of the key concepts are standardized across studies. Exploratory research, of course, is much less likely to use standardized measures. In such instances, concepts and measures are developed “along the way.” Such

research, however, should not be viewed as entirely unstructured and lacking standards of evaluation and interpretation. Maxwell (1996) convincingly shows that exploratory research can still be conducted within quasi-scientific guidelines.

A final measurement issue concerns the use of available data, which now is the major source of data for social science research. Often, these data were collected for some purpose other than what the researcher has in mind. In such instances, PROXY VARIABLES are often relied on, and validity issues usually must be confronted. Reliability is usually less of a problem because proxy measures often seem to be sociodemographic variables. Fortunately, large-scale data collection organizations are now relying more on input from social scientists for measures when they collect their data. Having to locate and use proxy measures of key concepts, it is hoped, will become a thing of the past.

—Charles W. Mueller

REFERENCES

- Bollen, K. (1989). *Structural equations with latent variables*. New York: John Wiley.
- DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park, CA: Sage.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Maxwell, J. A. (1996). *Qualitative research design*. Thousand Oaks, CA: Sage.

CONDITIONAL LIKELIHOOD RATIO TEST

One procedure for testing a restriction on parameters estimated with CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATION is the conditional likelihood ratio test. The intuition behind the conditional likelihood ratio statistic is analogous to that of the LIKELIHOOD RATIO STATISTIC. The conditional likelihood ratio statistic measures the reduction in the conditional log-likelihood function from imposing the restriction. If the restriction is valid, the reduction in the conditional log-likelihood function should be insignificant. The large sample distribution of the conditional likelihood ratio statistic is chi-squared with DEGREES OF FREEDOM equal to the number of restrictions imposed. As with the likelihood ratio statistic, the conditional likelihood ratio test has

counterparts in conditional versions of the Wald and score (Lagrange multiplier) tests.

More formally, suppose that a parameter vector, β , is estimated by maximizing the following conditional log-likelihood function, $\ln L = \sum f(y_2|\mathbf{X}, \beta, \mathbf{Z}, \gamma)$, where γ is a parameter vector whose values are set by theoretical assumption or, more commonly, replaced with the estimate, $\hat{\gamma}$, which is obtained by maximizing the marginal log-likelihood function, $\ln L = \sum g(y_1|\mathbf{Z}, \gamma)$. Now consider a $(r \times 1)$ vector of restrictions on β , defined as $\mathbf{c}(\beta) - \mathbf{q} = \mathbf{0}$. The conditional likelihood ratio test is based on the difference between $\ln \hat{L}$, the conditional log-likelihood function at the unrestricted estimate of β , and $\ln \hat{L}_R$, the conditional log-likelihood function at the restricted estimate of β . Note that both are conditioned on γ , which is treated as known. Then, the conditional likelihood ratio statistic is $-2(\ln \hat{L}_R - \ln \hat{L}) \sim \chi_r^2$, which tests $\mathbf{c}(\beta) - \mathbf{q} = \mathbf{0}$ as the null hypothesis.

One application of the conditional likelihood ratio test is the test of exogeneity in the context of simultaneous PROBIT models. Consider the specific example of a probit model in which one of the regressors, $y_1 = \mathbf{Z}\gamma + v$, is a continuous ENDOGENOUS variable with a NORMAL DISTRIBUTION. Conditional maximum likelihood estimation of this system of equations would involve obtaining a maximum likelihood estimate of γ by ORDINARY LEAST SQUARES and then using this estimate to condition the maximum likelihood estimation of the probit equation. The conditioning in this case is using $\hat{\gamma}$ to derive reduced-form residuals, $\hat{v}$, that would be included on the right-hand side of the probit equation to control for simultaneity bias. The conditional likelihood ratio test would compare the log-likelihood for this unrestricted model to the log-likelihood for the restricted model that imposes the restriction of exogeneity by excluding the reduced-form residuals. In this specific example, the conditional likelihood ratio statistic would have a chi-squared distribution with one degree of freedom. Note that under the null hypothesis of exogeneity, this conditional likelihood ratio test is asymptotically equivalent to the likelihood ratio test.

—Harvey D. Palmer

REFERENCES

- Alvarez, R. M., & Glasgow, G. (2000). Two-stage estimation of nonrecursive choice models. *Political Analysis*, 8, 147–165.
- Rivers, D., & Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39, 347–366.

CONDITIONAL LOGIT MODEL

Conditional logit is a statistical model used to study problems with unordered CATEGORICAL (NOMINAL) dependent variables that have three or more categories. Conditional logit estimates the probability that each observation will be in each category of the dependent variable. Conditional logit uses characteristics of the choice categories as independent variables and thus does not allow the effects of independent variables to vary across choice categories. This differs from the similar MULTINOMIAL LOGIT model, which uses characteristics of the individual observations as independent variables and allows the effects of the independent variables to vary across the categories of the dependent variable. Models that include both characteristics of the alternatives and characteristics of the individual observations are possible and are also usually referred to as conditional logits. Note that some researchers refer to both the multinomial logit and the conditional logit as a “multinomial logit.”

There are several different ways to derive the conditional logit model. The most common approach in the social sciences is to derive the conditional logit as a DISCRETE choice model. Each observation is an actor selecting one alternative from a choice set of J alternatives, where J is the number of categories in the dependent variable. The utility an actor i would get from selecting alternative k from this choice set is given by

$$U_{ik} = X_{ik}\beta + \varepsilon_{ik}.$$

Let y be the dependent variable with J unordered categories. Thus, the probability that an actor i would select alternative j over alternative k is

$$\begin{aligned} \Pr(y_i = k) &= \Pr(U_{ik} > U_{ij}) \\ &= \Pr(X_{ik}\beta + \varepsilon_{ik} > X_{ij}\beta + \varepsilon_{ij}) \\ &= \Pr(\varepsilon_{ik} - \varepsilon_{ij} > X_{ij}\beta - X_{ik}\beta). \end{aligned}$$

McFadden (1973) demonstrated that if the ε s are independent and there are identical Gumbel distributions across all observations and alternatives, this discrete choice derivation leads to a conditional logit. Multinomial logit estimates $\Pr(y_i = k|x_{ij})$, where k is one of the J categories in the dependent variable. The conditional logit model is written as

$$\Pr(y_i = k|x_{ik}) = \frac{e^{x_{ik}\beta}}{\sum_{j=1}^J e^{x_{ij}\beta}}.$$

Conditional logits are estimated using MAXIMUM LIKELIHOOD techniques. Unlike the multinomial logit, no identification constraints are necessary on those coefficients on variables that vary across alternatives as well as individuals. However, any coefficients on variables that do not vary across alternatives must be constrained in some way to identify the model. The most common constraint is to normalize the β s on the variables that do not vary across individuals to equal zero for one alternative.

INDEPENDENCE OF IRRELEVANT ALTERNATIVES

Conditional logits and multinomial logits have a property known as independence of irrelevant alternatives (IIA). For each observation, the ratio of the probabilities of being in any two categories is not affected by the addition or removal of other categories from the choice set. This is because the ratio of the choice probabilities for any two alternatives only depends on the variables and coefficients pertaining to those two alternatives. To see that this holds for conditional logit, observe the following:

$$\frac{\Pr(y_i = m|x_{im})}{\Pr(y_i = k|x_{ik})} = \frac{e^{x_{im}\beta} / \sum_{j=1}^J e^{x_{ij}\beta}}{e^{x_{ik}\beta} / \sum_{j=1}^J e^{x_{ij}\beta}} = \frac{e^{x_{im}\beta}}{e^{x_{ik}\beta}}.$$

Note that IIA is a property of the choice probabilities of the individual observations and is not a property that applies to the aggregate probabilities in the model. This property is often undesirable in many settings. The most commonly cited example of this is the *red bus/blue bus paradox*. Suppose that an individual has a choice of either commuting to work via car or via a bus that happens to be painted red. Let the probability this individual selects each alternative be 1/2. Now a second bus that happens to be blue is added to the choice set, identical to the red bus in every way except for color. Our intuition tells us that the choice probabilities should be 1/2 for the car and 1/4 for each bus. However, IIA necessitates that the choice probabilities for each mode of transportation be 1/3, to preserve the original ratio of choice probabilities between the car and the red bus. Note that the IIA property will hold for any discrete choice model that assumes the error term is distributed independently and identically (IID). The desire to relax the IIA property is one factor that has led to the increasing use of models such as the multinomial probit (MNP) in the study of multinomial choice problems in the social sciences.

AN EXAMPLE

As an example, consider the problem of estimating the effects of age, gender, income, and ideological distance on candidate vote choice in an election that has three candidates (A, B, and C). Age, gender, and income are variables that do not vary across alternatives and so are treated as they would be in a MULTINOMIAL LOGIT. That is, the coefficients on these variables will vary across candidates and must be constrained in some way to identify the model. In this case, the coefficients for these variables are constrained to zero for Candidate C. Ideological distance between an individual and a candidate is measured as the difference between an individual's self-placement on a 7-point ideological scale and his or her placement of the candidate on the scale. As this variable varies across both individuals and alternatives, there will be a single coefficient for the impact of ideological distance on candidate choice, and this coefficient does not need to be constrained in any way.

The results of estimating a conditional logit for this choice problem are presented in Table 1.

Table 1 Multinomial Logit Estimates, Multicandidate Election (coefficients for Candidate C normalized to zero)

<i>Independent Variables</i>	<i>Coefficient Value</i>	<i>Standard Error</i>
Ideological distance	−0.12*	0.04
<i>A/C Coefficients</i>		
Income	−0.51*	0.20
Age	−0.01	0.01
Female	0.08	0.15
Constant	0.52	0.39
<i>B/C Coefficients</i>		
Income	0.24*	0.09
Age	−0.06*	0.01
Female	0.27	0.23
Constant	0.43	0.67
Number of observations	1,000	

*Statistically significant at the 95% level.

Notice the coefficients on age, gender, and income vary across alternatives, and the coefficients for one alternative have been normalized to zero. This conditional logit tells us that as income increases, individuals are less likely to vote for Candidate A in comparison to Candidate C, but they are more likely to vote for Candidate B in comparison to Candidate C. As age increases,

individuals are less likely to vote for Candidate B in comparison to Candidate C. The coefficient on ideological distance does not vary across candidates and tells us that as the ideological distance between individuals and a candidate increases, individuals are less likely to vote for that candidate.

—Garrett Glasgow

REFERENCES

- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of econometrics* (pp. 105–142). New York: Academic Press.
- Train, K. (1986). *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*. Cambridge: MIT Press.

CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATION

An alternative to full-information maximum likelihood (FIML) estimation is conditional maximum likelihood estimation (CMLE), which simplifies the maximization problem by treating some of the parameters as known. CMLE involves the maximization of a conditional log-likelihood function, whereby the parameters treated as known are either fixed by theoretical assumption or, more commonly, replaced by estimates. Applications of CMLE range from the iterative Cochrane-Orcutt procedure for AUTOCORRELATION to two-step estimation procedures for RECURSIVE SIMULTANEOUS EQUATIONS. CMLE produces consistent estimates that are generally less efficient than limited-information maximum likelihood (LIML) estimates, even though the two methods are ASYMPTOTICALLY equivalent under some conditions. CMLE is used instead of FIML estimation when the full log-likelihood function is difficult to derive or maximize. Consequently, the application of CMLE has declined in frequency with advancements in optimization and the speed of computer computation. CMLE is also important, however, as an intermediate step in iterative methods that produce maximum likelihood estimates, such as the expectation-maximization (EM) algorithm employed in BAYESIAN analysis.

CMLE is still often applied as a means of estimating the parameters of a recursive system of

equations, such as simultaneous equations involving limited dependent variables as functions of continuous ENDOGENOUS regressors (e.g., simultaneous PROBIT and TOBIT models). Rather than maximizing the joint log-likelihood function for the system, the conditional log-likelihood function for the equation at the top of the system is maximized by replacing the parameters in equations lower in the system with maximum likelihood estimates. In this two-step procedure, the first step involves maximizing the marginal log-likelihood functions of endogenous regressors and then using the parameter estimates from this first step to condition the MAXIMUM LIKELIHOOD ESTIMATION of the parameters in the equation at the top of the system.

More formally, consider the joint distribution $h(y_1, y_2 | \mathbf{X}, \beta, \mathbf{Z}, \gamma, \lambda)$ of two RANDOM VARIABLES, y_1 and y_2 . Suppose that this joint distribution can be factored into a marginal distribution for y_1 , $g(y_1 | \mathbf{Z}, \gamma)$, and a conditional distribution for y_2 , $f(y_2 | \mathbf{X}, \beta, E[y_1 | \mathbf{Z}, \gamma], \lambda)$, where β is the parameter vector for the exogenous regressors in $\mathbf{X}$, and λ is the parameter that defines the relationship between y_2 and $E[y_1 | \mathbf{Z}, \gamma]$. CMLE of β and λ is a two-step process. The first step maximizes the marginal log-likelihood function, $\ln L = \sum g(y_1 | \mathbf{Z}, \gamma)$, to obtain the estimate, $\hat{\gamma}$. The second step maximizes the conditional log-likelihood function, $\ln L = \sum f(y_2 | \mathbf{X}, \beta, E[y_1 | \mathbf{Z}, \gamma], \lambda)$, replacing γ with $\hat{\gamma}$.

In the specific case of the simultaneous probit model, $g(\cdot)$ would be a normal density and $f(\cdot)$ would be a probit density. The first step would apply ORDINARY LEAST SQUARES to obtain the maximum likelihood estimate of γ . The second step would use this estimate to condition the maximum likelihood estimation of the probit equation. This conditioning would use $\hat{\gamma}$ to derive reduced-form residuals that would be included on the right-hand side of the probit equation to control for simultaneity bias. One attractive characteristic of this two-stage conditional maximum likelihood (2SCML) estimator is that the estimation results can be used to test for exogeneity (i.e., $\lambda = 0$). One version of this exogeneity test is a CONDITIONAL LIKELIHOOD RATIO TEST. MONTE CARLO evidence indicates that this 2SCML estimator performs favorably against two alternatives: INSTRUMENTAL VARIABLES probit and two-stage probit least squares (Alvarez & Glasgow, 2000; Rivers & Vuong, 1988). Also, under exogeneity, this 2SCML estimator is asymptotically equivalent to the LIML estimator.

—Harvey D. Palmer

REFERENCES

- Alvarez, R. M., & Glasgow, G. (2000). Two-stage estimation of nonrecursive choice models. *Political Analysis*, 8, 147–165.
- Blundell, R. W., & Smith, R. J. (1989). Estimation in a class of simultaneous equation limited dependent variable models. *Review of Economic Studies*, 56, 37–58.
- Greene, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Rivers, D., & Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39, 347–366.

CONFIDENCE INTERVAL

A $100(1 - \alpha)\%$ confidence interval is an interval estimate around a POPULATION parameter θ that, under repeated RANDOM SAMPLES of size N , would be expected to include θ 's true value $100(1 - \alpha)\%$ of the time. The confidence interval is a natural adjunct to PARAMETER ESTIMATION because it indicates the precision with which a POPULATION PARAMETER is estimated by a sample statistic.

The confidence level, $100(1 - \alpha)\%$, is chosen a priori. A TWO-SIDED confidence interval uses a lower limit L and upper limit U that each contain θ 's true value $100(1 - \alpha/2)\%$ of the time, so that together they contain θ 's true value $100(1 - \alpha)\%$ of the time. This interval often is written as $[L, U]$, and sometimes writers combine a confidence level and interval by writing $\Pr(L < \theta < U) = 1 - \alpha$. In some applications, a ONE-SIDED confidence interval is used. Confidence intervals may be computed for a large variety of population parameters such as the MEAN, VARIANCE, proportion, and R -SQUARED.

The confidence interval is closely related to the SIGNIFICANCE TEST because a $100(1 - \alpha)\%$ confidence interval includes all hypothetical values of the population parameter that cannot be rejected by a significance test using a significance criterion of α . In this respect, it provides more information than a significance test does. Confidence intervals become narrower with larger sample size and/or lower confidence levels.

The limits L and U are derived from a sample statistic (often the sample estimate of θ) and a SAMPLING DISTRIBUTION that specifies the probability of getting each value that the sample statistic can take. Thus, L and U also are sample statistics and will vary from one sample to another. Suppose, for example, that a standard IQ test has been administered to a random

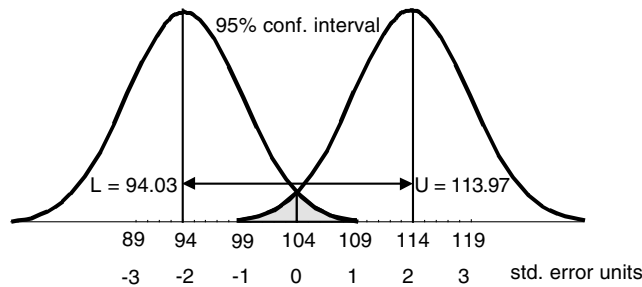


Figure 1 Lower and Upper Limits for the 95% Confidence Interval

sample of $N = 15$ adults from a large population with a sample mean of 104 and standard deviation (SD) of 18. We will construct a two-sided 95% confidence interval for the mean, μ . The limits U and L must have the property that, given a significance criterion of α , sample size of 15, mean of 104, and standard deviation of 18, we could reject the hypotheses that $\mu > 104 + U$ or $\mu < 104 - L$ but not $L < \mu < U$. Figure 1 displays those limits.

The sampling distribution of the T RATIO, where s_{err} stands for the standard error of the mean, defined by

$$t = \frac{\bar{X} - \mu}{s_{\text{err}}},$$

is a t -distribution with $df = N - 1 = 14$. When $df = 14$, the value $t_{\alpha/2} = 2.145$ standard error units above the mean cuts $\alpha/2 = .025$ from the upper tail of this t -distribution; likewise, $-t_{\alpha/2} = -2.145$ standard error units below the mean cuts $\alpha/2 = .025$ from the lower tail. The sample standard error is $s_{\text{err}} = \text{SD}/\sqrt{N} = 4.648$. So a t -distribution around $U = 104 + (2.145)(4.648) = 113.97$ has .025 of its tail below 104, while a t -distribution around $L = 104 - (2.145)(4.648) = 94.03$ has .025 of its tail above 104. These limits fulfill the above required property, so the 95% confidence interval for μ is [94.03, 113.97].

—Michael Smithson

REFERENCES

- Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). *Statistics with confidence: Confidence intervals and statistical guidelines* (2nd ed.). London: British Medical Journal Books.
- Smithson, M. (2003). *Confidence intervals* (Sage University Papers on Quantitative Applications in the Social Sciences, 140). Thousand Oaks, CA: Sage.

CONFIDENTIALITY

This ethical principle requires the researcher not to divulge any findings relating to research participants, other than in an anonymized form.

—Alan Bryman

See also ETHICAL PRINCIPLES

CONFIRMATORY FACTOR ANALYSIS

Confirmatory factor analysis is a statistical procedure for testing HYPOTHESES about the commonality among VARIABLES. As a MULTIVARIATE procedure, confirmatory factor analysis is used to simultaneously test multiple hypotheses that collectively constitute a measurement model. All such models comprise four components. At least two variables, or *indicators*, are necessary for the most rudimentary measurement model. Variability in scores on the indicators is allocated to two sources—LATENT VARIABLES and measurement ERROR. *Latent variables* are not observed; they are inferred from the commonality among indicators. *Measurement error* is variability in the indicators not attributable to the latent variables. Finally, the degree to which variability in scores on the indicators can be attributed to the latent variables is indexed by *loadings*, coefficients that are estimated in measurement equations in which the indicators are regressed on the latent variable(s) hypothesized to influence them.

SOURCES OF VARIABILITY

Confirmatory factor analysis partitions variability in scores on indicators according to sources specified in a measurement model. The basis for this partitioning is illustrated in Figure 1. On the left is a Venn diagram that shows six possible sources of variability in three indicators—V1, V2, and V3—that could be specified in a measurement model. In this informal depiction, the indicators are represented by circles, each comprising three parts. The shaded area, labeled F1, is variability in each indicator shared with the other two indicators. The crosshatched area is variability shared with only one of the remaining indicators, and the unshaded area is variability shared with neither of the remaining indicators.

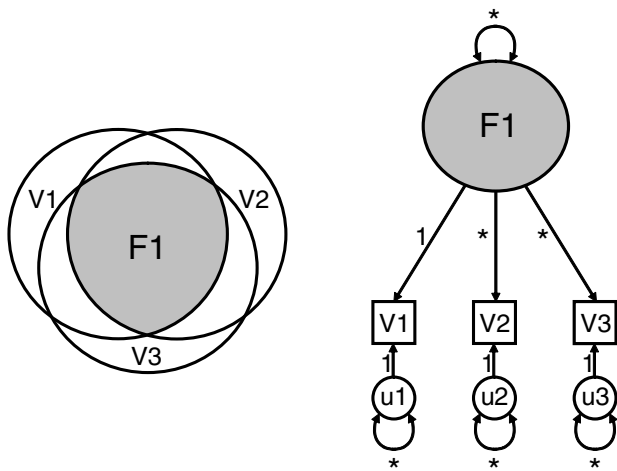


Figure 1 Venn and Path Diagram Depictions of the Partitioning of Variability in the Score on Three Variables Into Shared and Unique Components

On the right in Figure 1 is a more formal depiction of this partitioning of variability. In this **PATH DIAGRAM**, the indicators are represented by squares. The large shaded circle, labeled F1, corresponds to the shaded area in the Venn diagram. It is the latent variable the three indicators are purported to measure and reflects the variability they share. The directional *paths* between the latent variable and the squares represent the assumption that the latent variable is a cause of variability in the indicators. The small circles represent variability unique to each indicator and correspond to the unshaded areas in the Venn diagram. They represent measurement error because they are unintended causes of variability on the indicators. The crosshatched areas from the Venn diagram are not represented in the path diagram. This is because the typical specification of a measurement model assumes unrelated errors of measurement. If the strength of these relations is statistically nontrivial, a satisfactory measurement model will need to include covariances between the errors of measurement.

Two additional features of the path diagram warrant mention. First, notice that attached to each circle is a sharply curved two-headed arrow. This path indicates a **VARIANCE** and illustrates the fact that variability in each indicator has been allocated either to the latent variable or to measurement error. Second, notice that every path is accompanied by either a 1 or an asterisk. These represent **PARAMETERS**, numeric values that reflect the

strength of the paths. Parameters indicated by a value such as 1 reflect a fixed assumption about the association between two variables or a variance. Parameters indicated by an asterisk reflect a parameter whose value is not assumed but will be estimated from the observed data. A subset of these parameters of particular interest in confirmatory factor analysis is associated with paths between the latent variable and the indicators—the loadings. The larger these values, the greater the portion of variability in the indicators that can be attributed to the latent variable(s) as opposed to measurement error.

MODEL SPECIFICATION

Confirmatory factor analysis begins with the formal **SPECIFICATION** of a measurement model. That is, the researcher ventures a set of hypotheses in the form of sources and paths that account for the **ASSOCIATIONS** among the indicators. These hypotheses are evaluated against observed data on the indicators. To illustrate, imagine that a social psychologist wishes to evaluate the performance of a new measure of students' tendency to self-handicap: engaging in counterproductive behavior prior to important evaluative activities such as tests and presentations. She proposes two independent forms of this tendency and writes three items to tap each one. She then administers the measure to a sample of 506 college students.

The measurement model that reflects the researcher's hypotheses about the associations among the items is illustrated in Figure 2. The two latent variables, F1 and F2, represent the two hypothesized forms of self-handicapping. Note that each indicator, V1 to V6, is influenced by only one of the latent variables. Also note the curved arrow between F1 and F2. This path represents a **CORRELATION**. Because the researcher believes the two forms of self-handicapping are independent, it will be important for her to show that the value of this parameter, as estimated from her data is, in effect, zero.

It is customary in applications of confirmatory factor analysis to compare a proposed measurement model against plausible alternatives that are either more parsimonious than the proposed model or are consistent with rival theoretical accounts of the associations among the indicators. Our hypothetical researcher might compare her model against a more parsimonious model that includes only one latent variable or a plausible but less parsimonious one that includes two latent variables that

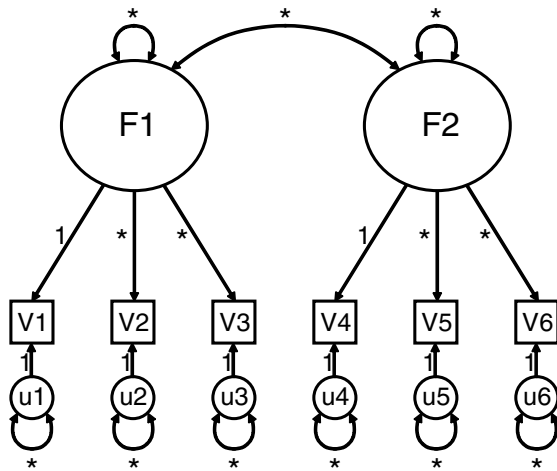


Figure 2 Path Diagram Depicting Two-Factor Model of Six-Item Measure of Self-Handicapping

are not independent. If, according to criteria presented below, the researcher's model provides a good account of the data in an absolute sense, and if it provides a better account than these two alternatives, then she has compelling evidence for the factorial VALIDITY of her measure.

Although confirmatory factor analysis allows considerable flexibility in the specification of measurement models, the mathematical requirements of estimation limit specification. For instance, a model in which each indicator in Figure 2 is influenced by both latent variables, the equivalent of an EXPLORATORY FACTOR ANALYSIS model, could not be estimated. A model in which none of the loadings was fixed at 1 could not be estimated without changing other aspects of the model. The principle that underlies these restrictions on specification is termed identification. Identification is a challenging topic that, fortunately, is rarely a concern in typical model specifications such as the one shown in Figure 2. Adherence to three basic rules will almost always result in a model that can be estimated. For each latent variable, there should be at least three indicators. The variance of, or one of the loadings on, each latent variable must be fixed at 1. In addition, there cannot be more unknown parameters in the model than the sum of the number of variances and covariances. In our example there are 6 variances, one for each indicator, and 15 covariances between the indicators, for a total of 21. (This number can be determined using the equation $p(p + 1)/2$, where p is the number of indicators.) Counting asterisks in the figure reveals a total of 13

unknown parameters. Our researcher's model meets the identification criteria and can be estimated.

PARAMETER ESTIMATION

Once a measurement model has been specified, it can be estimated. That is, values for unknown parameters are derived using data on the indicators. A number of ESTIMATORS could be used for this purpose, but MAXIMUM LIKELIHOOD ESTIMATION is virtually always the estimator of choice. The goal of maximum likelihood estimation is to determine the values of the unknown parameters that are most likely given the data. This determination is reached through an iterative process of substituting values for the unknown parameters, evaluating the likelihood of the parameters given the data, then attempting to improve on the likelihood by adjusting the values. Each evaluation is an *iteration*, and when it is not possible to improve significantly on the likelihood, the estimation process has *converged*. The parameter values at convergence are those on which hypothesis tests are based.

The iterative process cannot begin without a tentative value assigned to each unknown parameter. The simplest solution to this problem is to begin estimation with a value of 1 assigned to every unknown parameter. At the first iteration, these *starting values* will be adjusted, some upward and some downward, typically resulting in a substantial increase in the likelihood of the parameters given the data. After the first 2 or 3 iterations, improvement in the likelihood usually is slight, and convergence typically is reached in 10 or fewer iterations.

The parameter estimates for our example are shown in the top portion of Table 1. The estimates on which hypothesis tests are based are the UNSTANDARDIZED estimates. Notice that, among the loadings, V1, F1 and V4, F2 are 1, as prescribed in our model specification. Note that there are 15 entries under the "Parameter" heading. If we exclude the 2 loadings that were fixed, we are left with 13, each corresponding to one of the asterisks in Figure 2.

EVALUATION OF FIT

Having obtained the optimal set of parameter estimates for her model given the data she gathered, our researcher must now address two questions. First, does the model provide a satisfactory account of the data? Second, if the model provides a satisfactory account

Table 1 Parameter Estimates and Omnibus Fit Information

	Parameter Estimates					
Parameter	Unstandardized Estimate	Standard Error	Critical Ratio	Standardized Estimate		
Loadings						
V1, F1	1.000 ^a			.516		
V2, F1	1.236	.257	4.804	.437		
V3, F1	1.690	.380	4.445	.565		
V4, F2	1.000 ^a			.749		
V5, F2	.578	.106	5.472	.487		
V5, F2	.694	.127	5.472	.487		
Variances						
F1	.136	.038	3.572			
F2	.870	.175	4.690			
U1	.377	.039	9.594			
U2	.881	.073	12.034			
U3	.831	.104	8.027			
U4	.683	.158	4.326			
U5	.938	.078	12.042			
U6	1.347	.112	12.024			
Covariances						
F1, F2	−.004	.026	−.161	−.012		
Omnibus Fit Indices						
Model	χ^2	df	p	CFI	RMSEA	CL _(.90)
Two-factor model	13.820	8	.087	.976	.038	.000, .071
Drop F1, F2	13.845	9	.128	.980	.033	.000, .065
Drop F1, F2, add U1, U6	8.018	8	.432	1.000	.003	.000, .052
One-factor model	100.149	9	< .001	.622	.142	.117, .167

NOTE: CFI = comparative fit index; RMSEA = root mean square error of approximation. a. Fixed parameter.

of the data, do the parameter estimates conform to the predicted pattern? The first question concerns omnibus fit and the second component fit.

The *omnibus fit* of a measurement model is the degree to which the parameter estimates given the model imply a set of data that approximates the observed data. The word *data* in this case refers not to the individual responses of the 506 participants to the six self-report items. Rather, the data in question are elements of the observed covariance matrix. The observed covariance matrix for our example is shown in the top panel of Table 2. The matrix comprises 6 variances, one for each indicator, and 15 covariances, which index the pairwise associations between the indicators.

There are numerous inferential and descriptive statistics for indexing fit, each with its own rationale and interpretation. At the most basic level, however, the evaluation of omnibus fit is an exercise in comparison.

The most straightforward comparison is between the observed covariance matrix and the covariance matrix implied by the specified model given the parameters estimated from the data. To obtain the *implied covariance matrix*, one could substitute the estimated and fixed values of the parameters into the measurement equations and derive a covariance matrix. The implied covariance matrix for our example model is shown in the second panel of Table 2. The specified model fits this theoretical matrix exactly. To the extent that this matrix matches the observed covariance matrix, the specified model provides a satisfactory account of the observed data.

This comparison between the observed and implied covariance matrices is accomplished by subtracting each element of the implied covariance matrix from the corresponding element in the observed matrix, yielding the *residual matrix*. The residual matrix for our example is shown in the third panel of

Table 2 Observed, Implied, and Residual Covariance Matrices

<i>Observed Covariance Matrix</i>						
V1	0.513					
V2	0.169	1.090				
V3	0.231	0.284	1.221			
V4	0.031	−0.100	−0.066	1.553		
V5	0.035	−0.031	0.052	0.503	1.229	
V6	0.106	−0.017	−0.020	0.604	0.351	1.767
<i>Implied Covariance Matrix</i>						
V1	0.513					
V2	0.169	1.090				
V3	0.231	0.285	1.221			
V4	−0.004	−0.005	−0.007	1.553		
V5	−0.002	−0.003	−0.004	0.503	1.229	
V6	−0.003	−0.004	−0.005	0.604	0.350	1.767
<i>Residual Matrix</i>						
V1	0.000					
V2	0.000	0.000				
V3	0.000	−0.001	0.000			
V4	0.035	−0.095	−0.059	0.000		
V5	0.037	−0.028	0.056	0.000	0.000	
V6	0.109	−0.013	−0.015	0.000	0.001	0.000
<i>Standardized Residual Matrix</i>						
V1	0.000					
V2	0.000	0.000				
V3	0.000	−0.001	0.000			
V4	0.039	−0.073	−0.043	0.000		
V5	0.047	−0.024	0.046	0.000	0.000	
V6	0.114	−0.010	−0.010	0.000	0.001	0.000

Table 2. This matrix includes a number of zeros, which indicate no additional variance or covariance to be explained. The zeros on the diagonal are to be expected because all of the variance in the indicators is accounted for in the model either by the latent variables or measurement error. The zeros off the diagonal indicate that the researcher's model fully accounts for some associations between the indicators. Among the residuals that are not zero, some are positive and some are negative, indicating that the model underestimates some associations and overestimates others. Although the residual matrix is informative, particularly if respecification of the model is necessary, it does not provide

the basis for formally judging the degree to which the specified model provides a satisfactory account of the data.

A number of omnibus fit indices are available for this purpose. The most basic is the chi-square statistic, the basis for an inferential test of GOODNESS OF FIT. In our example, the value of chi-square evaluated on 8 degrees of freedom—the number of variances and covariances, 21, minus the number of free parameters, 13—is 13.82 with a *p* value of .09, indicating that the observed and implied covariance matrices are not significantly different. Two widely used indices are the comparative fit index (CFI) and the root mean square error of approximation (RMSEA). Values of the former range from 0 to 1, and values greater than .90 indicate satisfactory fit; the value of CFI in our example is .976. RMSEA originates at 0, and smaller values indicate better fit. Typically, a 90% confidence interval is put on the point estimate of RMSEA, and satisfactory fit is indicated by an upper limit less than .08. In our example, the upper limit is .071. This information in a format typical of reports of omnibus fit is shown in the first line of the lower portion of Table 1.

A second level of evaluation of fit concerns *component fit*, the degree to which the magnitude and sign of the parameter estimates correspond to predictions. In our example, we expect the factor loadings to be positive and significantly greater than zero. We expect the variances of the latent variables to be greater than zero as well. We venture no hypotheses regarding the measurement errors; however, smaller values indicate less variance in indicators attributable to measurement error. Finally, the researcher posited independent forms of self-handicapping; therefore, we predict that the covariance between F1 and F2 would not differ from zero.

Returning to the top portion of Table 1, we see that each of the 13 parameters estimated from the data is accompanied by a standard error. The ratio of the unstandardized parameter estimate to its standard error is termed the *critical ratio* and is distributed as a *z* statistic. As such, values greater than 1.96 indicate that the estimate is significantly different from zero. Only 1 of the 13 estimates is not significant, and it is the parameter our hypothetical researcher predicted to be zero: the F1, F2 covariance.

It is customary to interpret standardized factor loadings, which, because they are estimated in measurement equations, are STANDARDIZED REGRESSION COEFFICIENTS. A rule of thumb is that values greater

than .30 represent a significant contribution of the latent variable to the explanation of variability in the indicator. All six standardized loadings meet this criterion.

Our researcher's model provides a satisfactory account of her observed data at both the omnibus and component levels. To further strengthen this inference, she might compare this model against the two alternatives we described earlier. We established that the correlation between the latent variables is not different from zero; therefore, we know that a model in which this parameter is set to zero would fit the data as well as the initial model and, for reasons of parsimony, would be favored over it. The second alternative is a one-factor model. Omnibus fit indices for this model are presented in the last line of Table 1. Note that the value of chi-square is highly significant and substantially larger than the value for the two-factor model. The value of CFI is substantially less than .90, and both the upper and lower limits of the confidence interval exceed .08. Thus, there is strong support for our researcher's model in both an absolute and a comparative sense.

RESPECIFICATION

In practice, the initially specified measurement model does not yield satisfactory omnibus fit. In such instances, the researcher might attempt to *respecify* the model so as to find a defensible account of the data. At this point, the analysis shifts from hypothesis testing to hypothesis generating. That is, statistical tests are no longer, in a strict sense, valid. Nonetheless, respecification can be highly informative when evaluations of an initially specified model do not yield support for it.

Although the initially specified model in our example fit well, we might ask, for purposes of illustration, whether the fit can be improved. We already recognized that the model could be simplified by dropping the F1, F2 path, and the omnibus fit indices for such a model, shown in Table 1, support this assumption. Beyond this rather obvious adjustment to the model, it is not immediately evident how omnibus fit might be improved. In such cases, the researcher would resort to one of two *specification search* strategies. One uses one of the empirical strategies available in virtually all computer programs for confirmatory factor analysis. The other is a manual search strategy that focuses on the residual matrix. Because the elements of the residual matrix are covariances, which are not readily interpretable, we standardize them to

produce the standardized residual matrix shown in the bottom panel of Table 2. The values in this matrix are CORRELATION COEFFICIENTS and represent unexplained covariation between indicators. The largest of these is the correlation between V1 and V6. The simplest means of accounting for this association would be to allow the corresponding measurement errors, U1 and U6, to covary. When we respecify the model to include this parameter, the standardized residual drops to .024, and the omnibus fit indices, shown at the bottom of Table 1, approach their maxima. Were there no support for the model as initially specified, modifications such as this might be sufficient to achieve statistical support. It is important to keep in mind, however, that because such increments in fit are achieved using knowledge about the observed data, they might be sample specific and, therefore, must be confirmed through evaluation of the respecified model when specified a priori and estimated from new data.

ORIGIN AND CURRENT DIRECTIONS

Confirmatory factor analysis traces its roots to the work of Swedish psychometrician Karl Jöreskog. During the early 1970s, Jöreskog and associates developed algorithms and, importantly, computer software for analyzing the structure of covariance matrices. Their LISREL (*linear structural relations*) model allowed for the formal evaluation of COVARIANCE STRUCTURE MODELS, of which measurement models are a specific instance.

Most advances in confirmatory factor analysis are a by-product of advances in the general analysis of covariance structures. These include additional computer programs, some permitting model specification by sketching a path diagram; a host of omnibus fit indices that address documented shortcomings of the chi-square test; and estimation procedures that do not impose the demands on data that attend maximum likelihood estimation. Advances specific to confirmatory factor analysis typically manifest as solutions to specific research problems. Examples include tests of INTERACTION EFFECTS involving latent variables, latent GROWTH CURVE MODELS, and state-trait models.

—Rick H. Hoyle

REFERENCES

- Hoyle, R. H. (2000). Confirmatory factor analysis. In H. E. A. Tinsely & S. D. Brown (Eds.), *Handbook of*

- applied multivariate statistics and mathematical modeling* (pp. 465–497). New York: Academic Press.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183–202.
- Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

CONFOUNDING

The word *confounding* has been used to refer to at least three distinct concepts. In the oldest and most widespread usage, confounding is a source of bias in estimating causal effects. This bias is sometimes informally described as a mixing of effects of extraneous factors (called confounders) with the effect of interest. This usage predominates in nonexperimental research, especially in epidemiology and sociology. In a second and more recent usage originating in statistics, *confounding* is a synonym for a change in an effect measure on stratification or adjustment for extraneous factors (a phenomenon called *noncollapsibility* or *Simpson's paradox*). In a third usage, originating in the EXPERIMENTAL DESIGN literature, confounding refers to inseparability of main effects and interactions under a particular design. The three concepts are closely related and are not always distinguished from one another. In particular, the concepts of confounding as a bias in effect estimation and as noncollapsibility are often treated as equivalent, even though they are not. Only the former concept will be described here; for more detailed coverage and comparisons of concepts, see the following: Rothman and Greenland (1998, chaps. 4, 20); Greenland, Robins, and Pearl (1999); Pearl (2000); and Greenland and Brumback (2002).

CONFOUNDING AS A BIAS IN EFFECT ESTIMATION

A classic discussion of confounding in which explicit reference is made to “confounded effects” is Mill (1843/1956, chap. 10), although in Chapter 3, Mill lays out the primary issues and acknowledges Francis Bacon as a forerunner in dealing with them. There, he lists a requirement for an EXPERIMENT intended to determine causal relations: “None of the circumstances [of the experiment] that we do know shall have effects

susceptible of being *confounded* with those of the agents whose properties we wish to study” (emphasis added).

In Mill's time, the word *experiment* referred to an observation in which some circumstances were under the control of the observer, as it still is used in ordinary English, rather than to the notion of a comparative trial. Nonetheless, Mill's requirement suggests that a comparison is to be made between the outcome of our “experiment” (which is, essentially, an uncontrolled trial) and what we would expect the outcome to be if the agents we wish to study had been absent. If the outcome is not as one would expect in the absence of the study agents, then Mill's requirement ensures that the unexpected outcome was not brought about by extraneous “circumstances” (factors). If, however, those circumstances do bring about the unexpected outcome and that outcome is mistakenly attributed to effects of the study agents, then the mistake is one of confounding (or confusion) of the extraneous effects with the agent effects.

Much of the modern literature follows the same informal conceptualization given by Mill. Terminology is now more specific, with *treatment* used to refer to an agent administered by the investigator and *exposure* often used to denote an unmanipulated agent. The chief development beyond Mill is that the expectation for the outcome in the absence of the study exposure is now almost always explicitly derived from observation of a CONTROL GROUP that is untreated or unexposed. Confounding typically occurs when natural or social forces or personal preferences affect whether a person ends up in the treated or control group, and these forces or preferences also affect the outcome variable. Although such confounding is common in observational studies, it can also occur in randomized experiments when there are systematic improprieties in TREATMENT allocation, administration, and compliance. A further and somewhat controversial point is that confounding (as per Mill's original definition) can also occur in perfect randomized trials due to *random* differences between comparison groups (Fisher, 1935; Rothman, 1977).

THE POTENTIAL OUTCOME MODEL

Various models of confounding have been proposed for use in statistical analyses. Perhaps the one closest to Mill's concept is based on the *potential outcome* or COUNTERFACTUAL model for causal effects. Suppose we wish to consider how a health status (outcome)

measure of a population would change in response to an intervention (population treatment). More precisely, suppose our objective is to determine the effect that applying a treatment x_1 had or would have on an outcome measure μ relative to applying treatment x_0 to a specific target Population A. For example, Cohort A could be a cohort of breast cancer patients, treatment x_1 could be a new hormone therapy, x_0 could be a placebo therapy, and the measure μ could be the 5-year survival probability. The treatment x_1 is sometimes called the *index* treatment, and x_0 is sometimes called the *control* or *reference* treatment (which is often a standard or placebo treatment).

The potential outcome model posits that, in Population A, μ will equal μ_{A1} if x_1 is applied and μ_{A0} if x_0 is applied; the causal effect of x_1 relative to x_0 is defined as the change from μ_{A0} to μ_{A1} , which might be measured as $\mu_{A1} - \mu_{A0}$ or μ_{A1}/μ_{A0} . If A is given treatment x_1 , then μ will equal μ_{A1} and μ_{A1} will be observable, but μ_{A0} will be unobserved. Suppose, however, we expect μ_{A0} to equal μ_{B0} , where μ_{B0} is the value of the outcome μ observed or estimated for a Population B that was administered treatment x_0 . The latter population is sometimes called the control or reference population. Confounding is said to be present if $\mu_{A0} \neq \mu_{B0}$, for then there must be some difference between Populations A and B (other than treatment) that is affecting μ .

If confounding is present, a naive (crude) ASSOCIATION measure obtained by substituting μ_{B0} for μ_{A0} in an effect measure will not equal the effect measure, and the association measure is said to be *confounded*. For example, if $\mu_{A0} \neq \mu_{B0}$, then $\mu_{A1} - \mu_{B0}$, which measures the association of treatments with outcomes *across* the populations, is confounded for $\mu_{A1} - \mu_{A0}$, which measures the effect of treatment x_1 on Population A. Thus, to say an association measure $\mu_{A1} - \mu_{B0}$ is confounded for an effect measure $\mu_{A1} - \mu_{A0}$ is to say that these two measures are not equal.

A noteworthy aspect of this view is that confounding depends on the outcome measure. For example, suppose Populations A and B have a different 5-year survival probability μ under placebo treatment x_0 ; that is, suppose $\mu_{B0} \neq \mu_{A0}$, so that $\mu_{A1} - \mu_{B0}$ is confounded for the actual effect $\mu_{A1} - \mu_{A0}$ of treatment on 5-year survival. It is then still possible that 10-year survival, ν , under the placebo would be identical in both populations; that is, ν_{A0} could still equal ν_{B0} , so that $\nu_{A1} - \nu_{B0}$ is not confounded for the actual effect of treatment on 10-year survival. (We should generally

expect no confounding for 200-year survival because no treatment is likely to raise the 200-year survival probability of human patients above zero.)

A second noteworthy point is that confounding depends on the target population of INFERENCE. The preceding example, with A as the target, had different 5-year survivals μ_{A0} and μ_{B0} for A and B under placebo therapy, and hence $\mu_{A1} - \mu_{B0}$ was confounded for the effect $\mu_{A1} - \mu_{A0}$ of treatment on Population A. A lawyer or ethicist may also be interested in what effect the hormone treatment would have had on Population B. Writing μ_{B1} for the (unobserved) outcome under treatment, this effect on B may be measured by $\mu_{B1} - \mu_{B0}$. Substituting μ_{A1} for the unobserved μ_{B1} yields $\mu_{A1} - \mu_{B0}$. This measure of association is confounded for $\mu_{B1} - \mu_{B0}$ (the effect of treatment x_1 on 5-year survival in Population B) if and only if $\mu_{A1} \neq \mu_{B1}$. Thus, the same measure of association, $\mu_{A1} - \mu_{B0}$, may be confounded for the effect of treatment on neither, one, or both of Populations A and B and may or may not be confounded for the effect of treatment on other targets.

Confounders (Confounding Factors)

A third noteworthy aspect of the potential outcome model is that it invokes no explicit differences (imbalances) between Populations A and B with respect to circumstances or covariates that might influence μ (Greenland & Robins, 1986). Clearly, if μ_{A0} and μ_{B0} differ, then A and B must differ with respect to factors that influence μ . This observation has led some authors to define confounding as the presence of such covariate differences between the compared populations. Nonetheless, confounding is only a consequence of these covariate differences. In fact, A and B may differ profoundly with respect to covariates that influence μ , and yet confounding may be absent. In other words, a covariate difference between A and B is a necessary but not sufficient condition for confounding. For example, the confounding effects of covariate differences may balance each other out, leaving no confounding.

Suppose now that Populations A and B differ with respect to certain covariates and that these differences have led to confounding of an association measure for the effect measure of interest. The responsible covariates are then termed *confounders* of the association measure. In the above example, with $\mu_{A1} - \mu_{B0}$ confounded for the effect $\mu_{A1} - \mu_{A0}$, the factors responsible for the confounding (i.e., the factors that

led to $\mu_{A0} \neq \mu_{B0}$) are the confounders. It can be deduced that a variable cannot be a confounder unless it can affect the outcome parameter μ within treatment groups and it is distributed differently among the compared populations (e.g., see Yule, 1903, who however uses terms such as *fictitious association* rather than *confounding*). These two necessary conditions are sometimes offered together as a definition of a confounder.

Nonetheless, counterexamples show that the two conditions are not sufficient for a variable with more than two levels to be a confounder (Greenland et al., 1999).

PREVENTION OF CONFOUNDING

Perhaps the most obvious way to avoid confounding is estimating $\mu_{A1} - \mu_{A0}$ to obtain a reference Population B for which μ_{B0} is known to equal μ_{A0} . Such a population is sometimes said to be *comparable* to or *exchangeable* with A with respect to the outcome under the reference treatment. In practice, such a population may be difficult or impossible to find. Thus, an investigator may attempt to construct such a population or an exchangeable index and reference populations. These constructions may be viewed as design-based methods for the control of confounding.

Perhaps no approach is more effective for preventing confounding by a known factor than *restriction*. For example, gender imbalances cannot confound a study restricted to women. However, there are several drawbacks: Restriction on enough factors can reduce the number of available subjects to unacceptably low levels and may greatly reduce the generalizability of results as well. Matching the treatment populations on confounders overcomes these drawbacks and, if successful, can be as effective as restriction. For example, gender imbalances cannot confound a study in which the compared groups have identical proportions of women.

Unfortunately, differential losses to observation may undo the initial covariate balances produced by matching. Neither restriction nor matching prevents (although it may diminish) imbalances on unrestricted, unmatched, or unmeasured covariates. In contrast, *randomization* offers a means of dealing with confounding by covariates not accounted for by the design. It must be emphasized, however, that this solution is only probabilistic and subject to severe constraints in practice.

Randomization is not always feasible or ethical, and many practical problems, such as differential loss and noncompliance, can lead to confounding in comparisons of the groups actually receiving treatments x_1 and x_0 . One somewhat controversial solution to noncompliance problems is *intent-to-treat analysis*, which defines the comparison groups A and B by treatment assigned rather than treatment received. Confounding may, however, affect even intent-to-treat analyses, and (contrary to widespread misperceptions) the bias in those analyses can exaggerate the apparent treatment effect (Robins, 1998). For example, the assignments may not always be random, as when blinding is insufficient to prevent the treatment providers from violating the assignment protocol. Also, purely by bad luck, randomization may itself produce allocations with severe covariate imbalances between the groups (and consequent confounding), especially if the study size is small (Fisher, 1935; Rothman, 1977). *Blocked* (matched) randomization can help ensure that random imbalances on the blocking factors will not occur, but it does not guarantee balance of unblocked factors.

ADJUSTMENT FOR CONFOUNDING

Design-based methods are often infeasible or insufficient to prevent confounding. Thus, an enormous amount of work has been devoted to analytic adjustments for confounding. With a few exceptions, these methods are based on observed covariate distributions in the compared populations. Such methods can successfully control confounding only to the extent that enough confounders are adequately measured. Then, too, many methods employ parametric models at some stage, and their success may thus depend on the faithfulness of the model to reality. These issues cannot be covered in depth here, but a few basic points are worth noting.

The simplest and most widely trusted methods of adjustment begin with *stratification* on confounders. A covariate cannot be responsible for confounding within internally homogeneous strata of the covariate. For example, gender imbalances cannot confound observations within a stratum composed solely of women. More generally, comparisons within strata cannot be confounded by a covariate that is unassociated with treatment within strata. This is so whether or not the covariate was used to define the strata. Thus, one need not stratify on all confounders to control

confounding. Furthermore, if one has accurate background information on relations among the confounders, one may use this information to identify sets of covariates sufficient for adjustment (Pearl, 2000). Nonetheless, if the stratification on the confounders is too coarse (e.g., because categories are too broadly defined), stratification may fail to adjust for much of the confounding by the adjustment variables.

One of the most common adjustment approaches today is to enter suspected confounders into a model for the outcome parameter μ . For example, let μ be the mean (expectation) of an outcome variable of interest, Y ; let X be the treatment variable of interest; and let Z be a suspected confounder of the X – Y relation. Adjustment for Z is often made by fitting a GENERALIZED LINEAR MODEL $g(\mu) = \alpha + \beta x + \gamma z$ or some variant, where x and z are specific values of X and Z , and $g(\mu)$ is a strictly increasing function such as the natural log $\ln(\mu)$, as in LOG-LINEAR MODELING, or the logit function $\ln\{\mu/(1 - \mu)\}$, as in LOGISTIC REGRESSION; the estimate of β that results is then taken as the Z -adjusted estimate of the effect on $g(\mu)$ of a one-unit increase in X .

An oft-cited advantage of model-based adjustment methods is that they allow adjustment for more variables and in finer detail than stratification. If the form of the fitted model cannot adapt well to the true dependence of Y on X and Z , however, such model-based adjustments may fail to adjust for confounding by Z . For example, suppose Z is symmetrically distributed around zero within X levels, and the true dependence is $g(\mu) = g(\alpha + \beta x + \gamma z^2)$; then, using the model $g(\mu) = g(\alpha + \beta x + \gamma z)$ will produce little or no adjustment for Z . Similar failures can arise in adjustments based on *propensity scores*. Such failures can be minimized or avoided by using reasonably flexible models and by carefully checking each fitted model against the data.

Some controversy has occurred about adjustment for covariates in randomized trials. Although Fisher (1935, p. 49) asserted that randomized comparisons were “unbiased,” he also pointed out that they could be confounded in the sense used here. Resolution comes from noting that Fisher’s use of the word *unbiased* referred to the design and was not meant to guide analysis of a given trial. Once the trial is under way and the actual treatment allocation is completed, the unadjusted treatment effect estimate will be biased if the covariate is associated with treatment, and this bias can be removed by adjustment for the covariate (Greenland & Robins, 1986; Rothman, 1977).

Finally, it is important to note that if a variable used for adjustment is not a confounder, bias may be introduced by the adjustment (Pearl, 2000); this bias tends to be especially severe if the variable is affected by both the treatment and the outcome under study (Greenland, 2003).

—Sander Greenland

REFERENCES

- Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver & Boyd.
- Greenland, S. (2003). Quantifying biases in causal models. *Epidemiology*, 14, 300–306.
- Greenland, S., & Brumback, B. A. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology*, 31, 1030–1037.
- Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15, 413–419.
- Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14, 29–46.
- Mill, J. S. (1956). *A system of logic, ratiocinative and inductive*. London: Longmans, Green & Company. (Original work published 1843)
- Pearl, J. (2000). *Causality*. New York: Cambridge University Press.
- Robins, J. M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine*, 17, 269–302.
- Rothman, K. J. (1977). Epidemiologic methods in clinical trials. *Cancer*, 39, 1771–1775.
- Rothman, K. J., & Greenland, S. (1998). *Modern epidemiology* (2nd ed.). Philadelphia: Lippincott.
- Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, 2, 121–134.

CONSEQUENTIAL VALIDITY

Consequential VALIDITY emerged from concerns related to societal ramifications of assessments. The *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999) views validity as a unitary concept that measures the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose. Thus, consequential validity is that type of validity evidence that addresses the intended and unintended consequences of test interpretation and use (Messick, 1989, 1995).

Examples of harmful and unintended consequences that relate to the need to determine the consequential validity of an instrument include the following:

1. High-stakes testing may negatively affect present and future educational and employment opportunities when a single test score is used to make a decision.
2. Lower pass rates may occur on tests by individuals of certain subgroups, such as people with disabilities or racial/ethnic minorities, because of unequal opportunities or inappropriate demands of the testing situation in terms of written or spoken language.
3. A less diverse workforce may result because of employment tests that increase reluctance to seek employment or complete the required assessments.
4. Teachers may narrow the curriculum to teach only what will be on the test.

Messick (1995) suggests that researchers should distinguish between adverse consequences that stem from valid descriptions of individual and group differences from adverse consequences that derive from sources of test invalidity such as CONSTRUCT underrepresentation and construct-irrelevant variance. If differences are related to the latter, then test invalidity presents a measurement problem that needs to be investigated in the validation process. However, if the differences are the result of valid group differences, then the adverse consequences represent problems of social policy. It is the researcher's responsibility to ensure that low scores do not occur because the assessment is missing something relevant to that construct that, if it were present, would have permitted the affected persons to display their competence. In addition, causes of low scores because of irrelevant demands in the testing process can prevent an individual from demonstrating his or her competence.

The *Standards for Educational and Psychological Testing* also emphasize the need to distinguish between those aspects of consequences of test use that relate to the construct validity of a test and issues of social policy. Standard 1.24 addresses the issue of consequential validity:

When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is

intended to assess or to the test's failure fully to represent the intended construct. (American Educational Research Association, 1999, p. 23)

Group differences do not in themselves call into question the validity of a proposed interpretation; however, they do increase the need to investigate rival hypotheses that may be related to problems with the validity of the measurement.

—Donna M. Mertens

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Educational Measurement. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education.
- Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741–749.

CONSISTENCY. See PARAMETER ESTIMATION

CONSTANT

The term *constant* simply refers to something that is not variable. In statistics, responses are typically described as RANDOM VARIABLES, roughly meaning that the responses cannot be predicted with certainty. For example, how much weight will the typical adult lose following some particular diet? How will individuals respond to a new medication? If we randomly pick some individual, will he or she approve of the death penalty?

Although at some level, the difference between a constant and a random variable is clear, the distinction between the two often becomes blurred. Consider, for example, the population mean, μ . That is, μ is the average of all individuals of interest in a particular study if they could be measured. The so-called frequentist approach to statistical problems views μ as a constant. It is some fixed but unknown value. However, an alternative view, reflected by a Bayesian approach to statistics, does not view μ as a constant

but rather as a quantity that has some distribution. The distribution might reflect prior beliefs about the likelihood that μ has some particular value. As another example, p might represent the probability that an individual responds yes when asked if he or she is happily married. In some sense, this is a constant: At a particular moment in time, one could view p as fixed among all married couples. Simultaneously, p could be viewed as a random variable, either in the sense of prior beliefs held by the investigator or perhaps as varying over time.

Another general context in which the notion of constant plays a fundamental role has to do with assumptions made when analyzing data. Often, it is assumed that certain features of the data are constant to simplify technical issues. Perhaps the best-known example is HOMOSKEDASTICITY. This refers to the frequently made assumption that the VARIANCE among groups of individuals is constant. In REGRESSION, constant variance means that when trying to predict Y based on some variable X , the (conditional) variance of Y , given X , does not vary. So, for example, if X is aggression in the home and Y is a measure of cognitive functioning, constant variance means that the variance of Y among homes with an aggression score of, for example, $X = 10$ is the same as homes with $X = 15$, $X = 20$, and any other value for X we might choose.

—Rand R. Wilcox

REFERENCES

- Conover, W. J. (1980). *Practical nonparametric statistics*. New York: John Wiley.
- Doksum, K. A., & Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63, 421–434.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego: Academic Press.

CONSTANT COMPARISON

Constant comparison is the data-analytic process whereby each interpretation and finding is compared with existing findings as it emerges from the data analysis. It is associated with QUALITATIVE RESEARCH more than with QUANTITATIVE RESEARCH. It is normally associated with the GROUNDED THEORY data-analytic method, within which Glaser and Strauss (1967) referred to it as the “constant comparative method of qualitative analysis.” Qualitative and quantitative data can be subject to constant

comparison, but the analysis of those data is invariably qualitative.

Each comparison is usually called an iteration and is normally associated with INDUCTIVE reasoning rather than deductive reasoning; as a result, it is also referred to as “analytic induction” (Silverman, 1993). However, hypothetico-deductive reasoning will often occur within each iteration of the constant comparison method. Constant comparison is normally associated with the IDIOGRAPHIC philosophy and approach to research rather than the nomothetic philosophy and approach. Methodologies that normally employ constant comparison include ETHNOGRAPHY, PHENOMENOLOGY, SYMBOLIC INTERACTIONISM, and ETHNOMETHODOLOGY. Constant comparison contributes to the validity of research.

An example of constant comparison might be apparent when a researcher is researching the phenomenon of leadership within an organization. The methodology might employ interview data supported by observation and document data. The initial analysis of those data might involve CODING of interview transcripts to identify the variables, or categories, that seem to be present within the manifestation of the phenomenon. After analysis of the initial interviews, a number of categories might emerge, and relationships between those categories might be indicated. With each subsequent interview, each emerging category is compared with the extant categories to determine if the emerging category is a discrete category, a property of an existing category, or representative of a category at a higher level of abstraction.

For example, Kan (2002) used the full grounded theory method to research nursing leadership within a public hospital. She determined the presence of a number of lower order categories from the constant comparison of interviews and observations tracked over a 14-month period. In addition, Kan administered a questionnaire over this time frame to measure the leadership demonstrated by the managers in question. The comparison of the questionnaire results with the categories that emerged from observation and interview secured the identification of two higher order categories called *multiple realities* and *repressing leadership*.

This ongoing, or constant, comparison continues throughout the analysis of all data until the properties of all categories are clear and the relationships between categories are clear to the researcher. In the case of the work of Kan (2002), probing into the low reliabilities of some factors and the written

comments provided on the questionnaires provided insights that highlighted the characteristics of the higher order categories and ultimately confirmed the basic social process of “identifying paradox.”

—Ken W. Parry

REFERENCES

- Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Chicago: Aldine.
- Kan, M. (2002). Reinterpreting the multifactor leadership questionnaire. In K. W. Parry & J. R. Meindl (Eds.), *Grounding leadership theory and research: Issues, perspectives and methods* (pp. 159–173). Greenwich, CT: Information Age Publishing.
- Silverman, D. (1993). *Interpreting qualitative data: Methods for analysing talk, text, and interaction*. London: Sage.

CONSTRUCT

A construct is a concept used by social scientists to help explain empirical data on a phenomenon or to conceptualize unobservable or unmeasurable elements of a domain of study to formulate a theory. Specifically, it can mean (a) an idea or a concept not directly observable or measurable (i.e., a facet of a theory), (b) a concept inferred or constructed from empirical data to explain a phenomenon in an integrated way, or (c) an abstract (instead of operational) definition of a phenomenon or concept that is manifested in other observable or empirical concepts or consequences.

For example, logical reasoning is a construct that, according to some intelligence theories, represents a facet of intelligence. Logical reasoning is not directly observable; it is inferred from observable indices such as problem-solving ability, critical thinking ability, and others. Motivation is a construct that can be inferred from empirical data on learning; it helps to integrate diverse data to construct a theory of learning. Mathematical aptitude is an abstract definition of a behavior; its existence is inferred from such manifestations as mathematical achievement, mathematical anxiety, and the number of required or elective math courses a person ever completed.

In formulating personality and social intelligence theories, psychologists have used verbal and nonverbal judgmental data, provided by ordinary people, and MULTIDIMENSIONAL SCALING to identify and confirm constructs (Jackson, 2002). This approach was equally successfully applied to study clinical judgments of

psychopathology (Jackson, 2002). Still others (e.g., Digman, 1990) used FACTOR ANALYSIS to study personality dimensions such as *agreeableness*, *will*, *emotional stability*, and *intellect* (or openness). These constructs (i.e., factors) were extracted from CORRELATION or COVARIANCE matrices of numerous instruments for measuring personality. With advanced statistical inferential techniques such as structural equating modeling, Carroll (2002) showed that HIGHER-ORDER factors could be further extracted from these five factors to account for individual differences in personality and to help diagnose and treat personality disorders. In educational assessments, construct is a type of evidence that needs to be gathered to support the validity claim of an assessment instrument, according to the 1999 *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999). Messick also (1992) asserted that “the construct validity of score interpretation comes to undergird all score-based inferences—not just those related to interpretive meaningfulness, but including the content- and criterion-related inferences specific to applied decisions and actions based on test scores.” (p. 1493). For standards-based assessment programs, such as the National Assessment of Educational Progress (NAEP), measurement constructs are shaped by content standards and differential emphasis on different aspects of student understandings and performance.

For a construct to be useful for studying social phenomena, (a) it must be clearly defined, and (b) its relationships to similar or dissimilar concepts or constructs in the same domain of study must be specified. Together, the definition and the articulation of a construct’s relationship with other elements in a domain of study are sometimes referred to as “the theory of the construct” or its “nomological net” (Mueller, 1986). Even though a construct may be abstract, unobservable, or unmeasurable, its manifestation or consequence must be observable or measurable by some means for social scientists to study it, formulate theories, or test hypotheses about it.

—Chao-Ying Joanne Peng

See also CONSTRUCT VALIDITY

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

- Carroll, J. B. (2002). The five-factor personality model: How complete and satisfactory is it? In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 97–126). Mahwah, NJ: Lawrence Erlbaum.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41, 417–440.
- Jackson, D. N. (2002). The constructs in people's heads. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 3–18). Mahwah, NJ: Lawrence Erlbaum.
- Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (6th ed., p. 1493). New York: Macmillan.
- Mueller, D. J. (1986). *Measuring social attitudes: A handbook for researchers and practitioners*. New York: Teachers College, Columbia University.

CONSTRUCT VALIDITY

In educational assessment and social science measurement, construct validity is defined within the context of test validation. Test validation is the process whereby data are collected to establish the credibility of inferences to be made from test scores—both inferences about use of scores (as in decision making) and inferences about interpretation (meaning) of scores. Prior to the 1985 *Standards for Educational and Psychological Testing* (American Educational Research Association, 1985) validity was conceptualized, in the behaviorist tradition, as a taxonomy of three major and largely distinct categories (i.e., content validity, CRITERION RELATED validity, and construct validity). Construct validity is the extent to which the test is shown to measure a theoretical construct or trait. Constructs, such as creativity, honesty, and intelligence, are concepts used by social scientists to help explain empirical data about a phenomenon or to construct a theory. To execute the gathering of validity evidence, researchers must first define the construct in question. Theoretically based hypotheses (“validity hypotheses”) are then generated regarding the structure of the construct and its relationship to other constructs, as well as to any other human characteristics. Last, these validity hypotheses are supported or refuted with empirical data. A thorough validation requires the testing of many

validity hypotheses. Thus, test validation is a cumulative process.

Many techniques have been proposed to support construct validity, including FACTOR ANALYSIS, convergent and discriminant validation (multitrait-multimethod matrix proposed by Campbell & Fiske, 1959), examination of developmental changes (e.g., age differences on an intelligence test), internal consistency, experimental intervention, known-group differences, and correlation with other measures or tests that tap traits or behavior domains that are hypothesized to be related to the construct in question (Anastasi & Urbina, 1997; Cronbach & Meehl, 1955).

Loevinger (1957) argued that construct validity encompassed both content and criterion-related validity. This argument and others moved the 1985 edition and the most recent (1999) *Standards* to abandon the notion of three *types of validity*. Instead, validity is construed to be “a unified though faceted concept” (Messick, 1989) for which different *types of evidence* must be gathered to support inferences about the meaning of test scores and their proper use. It must be emphasized that inferences regarding the interpretation and use of the test scores are validated, not the test itself. Thus, construct validity is at the heart of the validation process—an evidence-gathering and inference-making progression. Professional organizations, such as the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), recommend the consideration of different *sources of validity evidence*, including, but not limited to, the following:

- test criterion evidence,
- evidence based on relations to other variables (such as correlation of test scores to external variables),
 - evidence based on test content,
 - convergent and discriminant evidence,
 - evidence based on internal structure (i.e., test items and test components conform to a theoretical construct), and
- generalizability (the extent to which the test criterion can be generalized to a new context without additional study of validity in the new context).

Messick (1989, 1992) proposed the matrix in Table 1, which summarizes key concepts in his conceptualization of the test validity model.

Table 1 Facets of Validity as a Progressive Matrix

	<i>Test Interpretation</i>	<i>Test Use</i>
Evidential basis	Construct validity (CV)	CV + relevance/utility (R/U)
Consequential basis	CV + value implications (VI)	CV + VI + R/U + social consequences

It should be noted that the field of research designs uses a slightly different nuance in its definition of construct validity. In this definition, the concern is with the veracity of the treatment/manipulation. Does the treatment effect a manipulation of the intended variable? Is the treatment true to the theoretical definition of the construct represented by an independent variable? In this context, the term is sometimes called the *construct validity of causes and effects* (Kirk, 1995).

—Chao-Ying Joanne Peng
and Daniel J. Mueller

See also CONSTRUCT, VALIDITY

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Prentice Hall.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod validity. *American Psychologist*, 15, 546–553.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(Suppl. 9), 635–694.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.
- Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (6th ed., p. 1487–1495). New York: Macmillan.

CONSTRUCTIONISM, SOCIAL

Social constructionism is, first of all, an account of knowledge-generating practices—both scientific and otherwise. At this level, constructionist theory offers an orientation toward knowledge making in the sciences, a standpoint at considerable variance with the empiricist tradition. At the same time, social constructionism contains the ingredients of a theory of human functioning; at this level, it offers an alternative to traditional views of individual, psychological processes. Constructionist premises have also been extended to a variety of practical domains, opening new departures in such fields as therapy, organizational management, and education. (For more complete accounts, see Gergen, 1994, 1999.) Of special relevance, they have contributed to the flourishing of many new forms of research methods in the social sciences.

SOCIAL CONSTRUCTIONIST ASSUMPTIONS

Social constructionism cannot be reduced to a fixed set of principles but is more properly considered a continuously unfolding conversation about the nature of knowledge and our understanding of the world. However, several themes are typically located in writings that identify themselves as constructionist. At the outset, it is typically assumed that our accounts of the world—scientific and otherwise—are not dictated or determined in any principled way by what there is. Rather, the terms in which the world is understood are generally held to be social artifacts, products of historically situated interchanges among people. Thus, the extent to which a given form of understanding prevails within a culture is not fundamentally dependent on the empirical validity of the perspective in question but rather on the vicissitudes of social process (e.g., communication, negotiation, communal conflict, rhetoric). This line of reasoning does not at all detract from the significance of various forms of cultural understanding, whether scientific or otherwise. People's constructions of the world and self are essential to the broader practices of a culture—justifying, sustaining, and transforming various forms of conduct. In addition, different communities of meaning making may contribute differentially to the resources available to humankind—whether it be “medical cures,” “moral intelligibilities,” institutions of law, or “reasons to live.” However, constructionism does challenge the warrant

of any group—science included—to proclaim “truth” beyond its perimeters. What is true, real, and good within one tradition may not be within another, and there are no criteria for judging among traditions that are themselves free of traditions, their values, goals, and way of life.

SOCIAL CONSTRUCTION AND SOCIAL SCIENCE

The social constructionist views favored by this composite of developments begin to furnish a replacement for traditional empiricist accounts of social science. In the process of this replacement, one may discriminate between two phases, deconstruction and reconstruction. In the former phase, pivotal assumptions of scientific rationality, along with bodies of empirically justified knowledge claims, are placed in question. This work essentially represents an elaboration and extension of the early anti-foundationalist arguments, now informed by the additional developments within the literary and critical domains. Thus, an extensive body of literature has emerged, questioning the warrant and the ideological implications of claims to truth, empirical hypothesis testing, universal rationality, laws of human functioning, the value neutrality of science, the exportation of Western scientific practices, and so on.

Immersion in this literature alone would lead to the conclusion that social constructionism is nihilistic in its aims. However, as many believe, the deconstructive process is only a necessary prolegomenon to a reconstructive enterprise. Within the reconstructive phase, the chief focus is on ways in which scientific inquiry, informed by constructionist views, can more effectively serve the society of which it is a part. From this emerging sensibility, several developments are noteworthy. First, constructionist ideas place a strong emphasis on theoretical creativity; rather than “mapping the world as it is,” the invitation is to create intelligibilities that may help us to build new futures. Theories of collaborative cognition, cyborg politics, and actor networks are illustrative. Second, constructionism has stimulated much work in cultural study, the critical and illuminating examination of everyday life practices and artifacts. Third, constructionist ideas have helped to generate a range of new practices in therapy, organizational change, and education in particular. Many scholars also find that in challenging disciplinary boundaries to knowledge, constructionist ideas invite

broad-ranging dialogue. Thus, new areas of interest have been spawned, linking for example, theology and constructionism, literary theory and social movements, and personality study and ethical theory.

SOCIAL CONSTRUCTION AND RESEARCH METHODS

Although much constructionist writing is critical of traditional empirical methods in the social sciences, these criticisms are not lethal. There is nothing about constructionist ideas that demands one kind of research method as opposed to another; every method has its ways of constructing the world. Thus, although traditional empiricist methods may be viewed as limited and ideologically problematic, they do have important uses. However, the major importance of constructionist ideas in the domain of methodology has been to incite discussion of new methods of inquiry. Although these new methods tend to be viewed as “qualitative” (Denzin & Lincoln, 2000), constructionists do not subscribe to the traditional qualitative/quantitative distinction that holds the former as preliminary and inferior to the latter. Most qualitative inquiry has different aims, different values, and a different politics than those inherent in quantitative inquiry. Thus far, constructionist work has functioned, in particular, to support research methods emphasizing the following:

- *Value reflection:* Who is advantaged by the research methods and who may be discredited? Is the research subject exploited by the research or treated as a mere object?
- *Subject voice:* Is the voice of the subject of research heard or legitimated by the method or obliterated by the research procedure?
- *Collaborative participation:* Can the subjects of research participate with the researcher in the generation of knowledge? Can they share in or benefit from the outcomes?
- *Multiple standpoints:* Are multiple viewpoints and values represented in the research, or does one standpoint dominate?
- *Representational creativity:* Must the representation of research be limited to formal writing, or can more populist and richly compelling forms of representation be located?

In its research emphases, constructionist assumptions are particularly evident in PARTICIPATORY ACTION

RESEARCH (Reason & Bradbury, 2000), discourse analysis (Wetherell, Taylor, & Yates, 2001), narrative inquiry (see, e.g., Josselsyn, 1996), participatory ethnography, and literary and performative approaches to representation (see, e.g., Ellis & Bochner, 1996).

—Kenneth J. Gergen

REFERENCES

- Denzin, N. K., & Lincoln, Y. S. (Eds.). (2000). *Handbook of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.
- Ellis, C., & Bochner, A. P. (Eds.). (1996). *Composing ethnography*. Walnut Creek, CA: AltaMira.
- Gergen, K. J. (1994). *Realities and relationships*. Cambridge, MA: Harvard University Press.
- Gergen, K. J. (1999). *An invitation to social construction*. London: Sage.
- Josselsyn, R. (Ed.). (1996). *Ethics and process in the narrative study of lives*. Thousand Oaks, CA: Sage.
- Reason, P., & Bradbury, H. (Eds.). (2000). *Handbook of action research, participative inquiry and practice*. Sage: London.
- Wetherell, M., Taylor, S., & Yates, S. J. (Eds.). (2001). *Discourse theory and practice*. London: Sage.

CONSTRUCTIVISM

Although found in various forms throughout philosophy and science, constructivism emerged in psychology as a theory of learning. It was based on Jean Piaget's (1929, 1937/1955) research with children, which indicated that children actively construct their own knowledge out of their experience rather than simply absorb what they are told by teachers. Jerome Bruner (1960, 1966) later took up the theory and applied it to education. He recommended that teaching should be designed in such a way that it encourages children to reflect on and use their existing knowledge and capabilities. In this way, they come to understand new concepts by working things out for themselves.

However, constructivism is also used more broadly to refer to a number of theories and approaches in psychology that see the person as actively engaged in the construction of his or her own subjective world. This is in opposition to views that regard objects and events as having an essential nature and universal meaning, where perception is a matter of internalizing a truthful representation of the world. Ernst von Glasersfeld (1984) coined the term *radical constructivism*,

arguing that all our knowledge of the world is constructed rather than gained through perception. We are therefore the constructors of our own subjective realities, the world that we each personally inhabit. For example, it is commonplace for those who have witnessed the same event to have quite different memories and descriptions of it. Radical constructivism would say that this is not because perception and memory are unreliable but because each person has constructed the event in a different way. Radical constructivism makes no claim about the existence or nature of an objective reality outside of our constructions.

A similar position is espoused by George Kelly, the founder of personal construct psychology (PCP). Kelly (1955) argues that each of us develops a system of dimensions of meaning, called constructs. We perceive the world in terms of these constructs, and our conduct can be understood in light of our own idiosyncratic construal of the world. Kelly, a clinical psychologist, devised a flexible method for exploring his clients' constructions that he called the REPERTORY GRID TECHNIQUE. This has since been adapted as a research tool in a diversity of settings outside the clinical field.

Constructivism shares some basic assumptions with social constructionism but differs from it in the extent to which the individual is seen as an agent who is in control of the construction process and in the extent to which our constructions are the product of social forces. Chiari and Nuzzo (1996) have attempted to provide a basis on which to differentiate the various constructivisms and suggest a distinction between epistemological and hermeneutic constructivism. Within hermeneutic constructivism, which would include social constructionism, reality is constituted through language and discourse. Epistemological constructivism, which would include PCP, emphasizes the construction of our personal worlds through the ordering and organization of our experience.

—Vivien Burr

REFERENCES

- Bruner, J. (1960). *The process of education*. Cambridge, MA: Harvard University Press.
- Bruner, J. (1966). *Toward a theory of instruction*. Cambridge, MA: Harvard University Press.
- Chiari, G., & Nuzzo, M. L. (1996). Psychological constructivisms: A metatheoretical differentiation. *Journal of Constructivist Psychology*, 9, 163–184.

- Kelly, G. (1955). *The psychology of personal constructs*. New York: Norton.
- Piaget, J. (1929). *The child's conception of the world*. New York: Harcourt Brace Jovanovich.
- Piaget, J. (1955). *The construction of reality in the child*. New York: Routledge Kegan Paul. (Original work published 1937)
- von Glasersfeld, E. (1984). An introduction to radical constructivism. In P. Watzlawick (Ed.), *The invented reality*. New York: Norton.

CONTENT ANALYSIS

Among the earliest definitions of content analysis are the following: "The technique known as content analysis . . . attempts to characterize the meanings in a given body of discourse in a systematic and quantitative fashion" (Kaplan, 1943, p. 230). "Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication" (Berelson, 1954, p. 489). Later definitions abandoned the requirement of quantification. Emphasis shifted to "inference," "objectivity," and "systematization." Holsti, in one of the early textbooks, wrote, "Content analysis is any technique for making inferences by objectively and systematically identifying specified characteristics of messages. . . . Our definition does not include any reference to quantification" (Holsti, 1969, p. 14), a point later reiterated by other contributors to the development of the technique (e.g., Krippendorff, 1980). Whether quantitative or simply inferential, content analysis approaches the study of text as a scientific, rather than simply interpretive, activity.

A COLLECTION OF DIFFERENT APPROACHES

What goes under the common label of content analysis is not a single technique; rather, it is a collection of different approaches to the analysis of texts or, more generally, of messages of any kind—from the word counts of the simplest forms of syntactical analysis to thematic analysis, referential analysis, and prepositional analysis (Krippendorff, 1980, pp. 60–63). To this list we must add a more recent approach based on story grammars and concerned with the study of social relations.

Word Counts and Key Word Counts in Context

The advance of computers has made computing the frequencies of the occurrence of individual words and concordances easy (i.e., words in the context of other words or key word in context [KWIC]) (Weber, 1990, pp. 44–53). This type of analysis (also known as corpus analysis among linguists) can provide a valuable first step in the process of data analysis of texts (using such corpus software as WordSmith).

Thematic Content Analysis

Thematic analysis is the most common approach in content analysis. In thematic analysis, the coding scheme is based on categories designed to capture the dominant themes in a text. Unfortunately, there cannot be a universal coding scheme: Different texts emphasize different things, and different investigators could be looking for different things in the same texts. That is why the development of good thematic analysis requires intimate familiarity with the input text and its characteristics. It also requires extensive pretesting of the coding scheme.

Referential Content Analysis

Thematic content analysis is a good tool for teasing out the main themes expressed in a text. But meaning is also the result of other kinds of language games: backgrounding and foregrounding information, silence and emphasis, or different ways of describing the same thing. Referential content analysis is a tool better suited than thematic analysis to capture the complexity of language in the production of meaning. Referential analysis, Krippendorff (1980) tells us, is used when "the task is to ascertain how an existing phenomenon is portrayed" (p. 62). That portrayal is affected by the choice of nouns and adjectives or even of different syntactical constructs, such as passive or active forms (Franzosi, 2003).

"Story Grammars" and the Structure of Narrative

From the very early days of the technique, researchers have pleaded for an approach to content analysis grounded on the linguistic properties of the text (e.g., de Sola Pool, Hays, Markoff, Shapiro, and Weitman; see Franzosi, 2003, chap. 1). In recent

years, several authors have attempted to bridge the gap between linguistics and content analysis, proposing some basic variants of “story grammars” (or text grammars or semantic grammars, as they are also known) for the analysis of narrative texts (Abell, 1987; Franzosi, 2003; Shapiro & Markoff, 1998). Basically, a story grammar is nothing but the simple structure of the five Ws of journalism: who, what, where, when, and why (and how); someone doing something, for or against someone else, in time and space; or subject-action-object or, better, agent-action-patient/beneficiary and their modifiers (e.g., type and number of subjects and objects, time and space, reasons and outcomes of actions). The technique has been used to great effect in the study of social and protest movements.

HOW CONTENT ANALYSIS WORKS

Quantification may not necessarily be part of any formal definition of content analysis, but what the different approaches to content analysis have in common is a concern with numbers—which distinguishes them from typically qualitative approaches to the analysis of texts and symbolic material, such as NARRATIVE ANALYSIS, CONVERSATION ANALYSIS, DISCOURSE ANALYSIS, or SEMIOTICS. But, if so, how does content analysis transform words into numbers? The engine of this transformation is certainly the “coding scheme,” although other “scientific” procedures play a vital role in the process: SAMPLING, VALIDITY, and RELIABILITY.

Coding Scheme

The coding scheme is the set of all coding categories applied to a collection of texts, in which a “coding category” identifies each characteristic of interest to an investigator. The scheme is systematically applied to all selected texts for the purpose of extracting uniform and standardized data: If a text contains information on any of the coding categories of the coding scheme, the relevant coding category is “ticked off” by a human coder (a process known as CODING in content analysis; modern qualitative software such as NUD*IST/N6 or Atlas.ti allow users to code text directly on the computer and to make up coding categories as they go along). When coding is completed, the ticks are added up for each coding category. In the end, the alchemic recipe of content analysis for getting numbers out of words is simple: You count. Whether it is words, themes, references, or actors and

their actions, depending on the specific type of content analysis, the numbers are the result of counting.

Sampling

Sampling provides an efficient and cost-effective way to achieve research results. Rather than working with all the possible available sources and documents, investigators may decide to sample an appropriate number of sources and an appropriate number of documents. Sampling may be particularly appealing in content analysis given that it is a very labor-intensive and, therefore, expensive technique (and sampling can also be used to check the reliability of coded data through acceptance sampling schemes) (Franzosi, 2003). By sampling, you may focus on a subset of newspapers, letters, or other documents of interest. You may focus on selected years. Even for those years, you may read one or more days a week of a newspaper, one or more weeks a month, or one or more months a year. Setting up a sampling frame that will not bias the data requires a sound knowledge of the input material. For example, how would relying on only the Sunday issues of a newspaper affect data? To answer this and similar questions, you should base any choice of sources and sampling frame on systematic comparative analyses. The validity of your data (i.e., the basic correspondence between a concept that you are trying to measure and your actual measurements) will be a function of the sources and sampling frame adopted.

Intercoder Reliability

In content analysis, issues of reliability (i.e., of the repeatability of measurement) are known as *intercoder reliability*. Would different coders, confronted with the same text and using the same coding scheme, “tick off” the same coding categories? The answer to that question partly depends on the design of the coding scheme and coding categories: The more abstract and theoretically defined the categories, the more likely that different coders will come up with different results. It is good practice to test the reliability of each coding category by having different coders code the same material.

AN EXAMPLE

Content analysis can be applied to a variety of texts (e.g., newspaper editorials, speeches, documents,

letters, ethnographic field notes, transcripts from in-depth interviews or focus groups) or images (e.g., advertisements, photographs). Below is a transcript of a focus group of young British-born Pakistani women who married men from Pakistan. The participants are as follows: Nadia (N), Laila (L), Ayesha (A), and the moderator (M: Maria Zubair, to whom I am grateful for the transcripts).

M: How did your husbands find life here—when they got here?

L: I think it depends really on how they live . . . when they're in Pakistan. I mean some people, you know—they're less fortunate than others, they live in villages or whatever, you know, they don't really go to college, but then there are those who live in towns and they have decent upbringing—you know so when they come here they don't—they are obviously—they're quite different but not so much. . . .

L: Yeah, I mean with my husband . . . when he came over here, he was the only child—um—he'd been to college, university and everything so he's educated—so when he came here he didn't find it too much of a difference. . . . So of course they find it different because it's . . . a lot different, isn't it!

L: You know going out—they find it difficult to make friends, I think that's the main thing.

A: Yeah.

L: To make friends—or places to go—you know sometimes in the evenings and weekends we say oh where shall we go, but there's nowhere to go you know—clubs and pubs they're for English people they're not for us.

L: They find it difficult anyway because for them to come and mix with—um British-born Asians is—it's a lot difficult—it is more difficult because they're just not on the same wavelength, you know, I mean my husband—he thinks Pakistaniwise and the boys from here they think differently—don't they!

A: I find that—I find that with my husband as well.

L: Yeah.

A: Sometimes I just sit there and um—we don't (laughs)—we're not on the same wavelength . . . he sees things the Pakistani way. . . . The time he comes home, he just sits there in front

of it [TV]—on sofa—right there. He won't move then. He just sits and that's it, you know. (laughs)

L: My husband used to say to me when he came over here “oh I don't like it” . . . “oh I don't like you wearing English clothes and stuff—oh why don't you wear hijab [head scarf],” and you know, all that stuff. “Why can't you wear Asian clothes to work.” . . . now he never says that, you know (N: yeah). So he's adapted to it . . . but when they first come here they do get a cultural shock—I mean everything is different (M: yeah) from there, you know, so—over there it is, isn't it. It's totally different.

A: (laughs) My husband—he, you know, he just winds me up. He used to wear shalwar qameez [Pakistani clothes] when he first got here—and I said to him, look when you go out . . . you're not going in these.

Even a simple word count reveals the dichotomous nature of the social context (perhaps, not unsurprisingly, given the moderator's leading question). The words occurring most frequently are *they* (21), *he* (16), *I* (12), *here* (9), and *there* (6). The relative high frequency of such words as *different* (5) and *difficult* (5) also goes a long way in highlighting the complex social relations of a marriage between a British-born Pakistani woman and a Pakistani man. After all, such words as *same* and *easy* never appear in the text. Thematic coding categories would also help to tease out the peculiarities of these different worlds. Whether done with the help of a specialized computer software or simply on paper, a coding scheme based on thematic categories would have to tap the following spheres of social life: family, friendships, jobs, and leisure. But these categories are too broad. The coding scheme would have to zoom in on the “difficulty/easiness” in “marital relationships,” “own family and in-laws,” “male and female friends,” and “friends in Pakistan and in England.” Frequency distributions of these categories would offer further insights on the question, “How did your husbands find life here?”

Yet there is more to the meaning of a text. Take Laila's words: “some people . . . less fortunate than others.” The choice of words is not random. Consider the alternatives: “less brave,” “less driven,” or “less adventurous.” No—they are simply “less fortunate.” There is no blame on them. Just the luck of the draw.

Conservative meritocratic views of the world would have us believe that we are all dealt a fair deck of cards. What we do with them is up to us, to the brave, the daring, the entrepreneurial. Implicitly, Laila frames her discourse along more liberal rather than conservative lines. But she also frames her discourse on the basis of implicit social scientific models that we can easily recognize as such. People who grow up in villages do not go to college. People in town have a decent upbringing. They get an education. They are not so different. Social scientific models of urbanization and of the relation between education and tolerance lurk behind Laila's words. Finally, Laila frames her discourse in very dichotomous terms of "us" versus "them": "clubs and pubs they're for English people they're not for us," "he thinks Pakistaniwise," and "we are not on the same wavelength." Laila and the other women in the group, as well as their husbands, are different from the English. But they are also different from their husbands. And Laila's educated husband is different from some of the other husbands.

These different modes of argumentation and framing could be highlighted through such coding categories as the following: "Conservative Political Views" and "Liberal Political Views," "Inclusive Worldviews" and "Exclusive/Dichotomous Worldviews," and "Mode of Reasoning (or Argumentation)," with such subcategories as "Emotive" and "Social Scientific." An emphasis on rhetoric and modes of argumentation would lead us to include such categories as "Appeal to Religion," "Appeal to Science," "Appeal to Emotion," "Appeal to Reason," "Appeal to Values of Equality," and so forth.

Referential analysis could be used to tap the differential portrayal of social actors and situations. For each main actor (e.g., the English, the husbands, the wives, the friends, the relatives back in Pakistan), event (e.g., visiting Pakistan, getting married, moving to England), or situations (e.g., an evening at home, time spent on weekends, dressing and eating habits), we could have three basic coding referential categories: positive, negative, or neutral.

CONCLUSIONS

Content analysis, as a typically quantitative approach to the study of texts, offers invaluable tools for teasing out meaning from texts or any other symbolic material. If confronted with the analysis of hundreds of pages of transcripts of the kind illustrated

above (or of any other types of documents/symbolic material), the variety of content analysis techniques would certainly help us reveal patterns in the data. One advantage of quantitative analysis is that it is efficient: It helps to deal with large quantities of data without getting buried under the sheer volume of material. Its disadvantage is that it may miss out on subtle nuances in the production of meaning. Even the few examples discussed above should have made that very clear. To think in terms of a quantitative versus a qualitative approach to texts is a misguided approach. Each has strengths and weaknesses. (Consider this: How long would it take to do a qualitative analysis of hundreds of pages of transcripts? How would you ensure that you are not relying on the best sound bytes or that you have not forgotten anything?) Theoretical/methodological developments in the analysis of texts coupled with technological developments in computer software are increasingly blurring the lines between quantitative and qualitative approaches (e.g., computer-aided qualitative data analysis software [CAQDAS]) (Fielding & Lee, 1998; Kelle, 1995). At the current state of the art, the best approach is probably one that combines a broad picture of a social phenomenon through quantitative content analysis with sporadic and deep incursions into selected texts using more qualitative approaches.

—Roberto Franzosi

REFERENCES

- Abell, P. (1987). *The syntax of social life: The theory and method of comparative narratives*. Oxford, UK: Clarendon.
- Berelson, B. (1954). Content analysis. In G. Lindzey (Ed.), *Handbook of social psychology* (Vol. 1, pp. 488–522). Reading, MA: Addison-Wesley.
- Fielding, N., & Lee, R. (1998). *Computer-analysis and qualitative research*. London: Sage.
- Franzosi, R. (2003). *From words to numbers*. Cambridge, UK: Cambridge University Press.
- Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.
- Kaplan, A. (1943). Content analysis and the theory of signs. *Philosophy of Science*, 10, 230–247.
- Kelle, U. (Ed.). (1995). *Computer-aided qualitative data analysis*. London: Sage.
- Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage.
- Shapiro, G., & Markoff, J. (1998). *Revolutionary demands: A content analysis of the Cahier de Doléances of 1789*. Stanford, CA: Stanford University Press.
- Weber, R. P. (1990). *Basic content analysis*. Newbury Park, CA: Sage.

CONTEXT EFFECTS. See ORDER EFFECTS

CONTEXTUAL EFFECTS

The notion of groups plays a major role in social science. Does membership of an individual in a particular group have any impact on thoughts or actions of the individual? Does living in a neighborhood consisting of many Democrats mean that an individual has a higher probability of being a Democrat than otherwise would have been the case? Such an impact by the group is known as a *contextual effect*. The study of these effects is called contextual analysis or MULTILEVEL ANALYSIS. It is also known as the study of HIERARCHICAL models.

If contextual effects exist, it raises the following questions: Why does belonging to a group have an impact, how does this mechanism work, and how should we measure the impact? The first two questions may best be left to psychologists, but the third question has sparked much interest among statisticians and others. In particular, in education, a classroom makes an obvious group for contextual analyses.

When we have data on two variables on individuals from several groups, within each group we can analyze the relationship of a dependent variable Y and an independent variable X . Regression analysis gives a SLOPE and an INTERCEPT for the regression line for each group. It may now be that the slopes and intercepts vary from one group to the next. If this variation in the regression lines is larger than can be explained by randomness alone, then there must be something about the groups themselves that affects the lines.

There can be several reasons why the slopes and intercepts vary. It may be that both depend on the level of the X variable in the group through the group mean. Thus, one possible contextual model can be expressed in the following equations:

$$\begin{aligned} a_j &= c_0 + c_2\bar{x}_j + u_j, \\ b_j &= c_1 + c_3\bar{x}_j + v_j, \end{aligned}$$

where a and b are the intercept and slope in the j th group, and the u and v are residual terms.

It is also possible to substitute for a and b in the equation for the line in the j th group. That gives the following equation:

$$y_{ij} = c_0 + c_1x_{ij} + c_2\bar{x}_j + c_3(x_{ij}\bar{x}_j) + e_{ij}$$

for the i th individual in the j th group. Thus, we can now run a MULTIPLE REGRESSION on data from all the groups and get the c coefficients. For this model, c_1 measures the effect of X on the level of the individual, c_2 measures the effect of X on the level of the group, and c_3 measures the effect on Y of the interaction of the individual and the group-level effects. One problem with this analysis is that the RESIDUAL E depends on the group means of X ; another is the COLLINEARITY that exists in the three explanatory variables.

Contextual analysis is not limited to the case in which the group intercepts and slopes depend on the group means of X . They could also depend on a variety of other variables.

Bayesian statistics lends itself particularly well to multilevel analysis because the slopes and intercepts vary from group to group. It may then be possible to specify how the intercepts and slopes within the groups come from specific distributions.

—Gudmund R. Iversen

REFERENCES

- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.
- Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffin.
- Iversen, G. R. (1991). *Contextual analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–081). Newbury Park, CA: Sage.

CONTINGENCY COEFFICIENT. See SYMMETRIC MEASURES

CONTINGENCY TABLE

A contingency table is a statistical table classifying observed data frequencies according to the categories in two or more variables. A table formed by the cross-classification of two variables is called a *two-way contingency table* (for an example, see

Table 1), a table of three variables is termed a *three-way contingency table*, and, in general, a table of three or more variables is known as a *multiway contingency table*. The analysis of multiway contingency tables is sometimes known as multivariate contingency analysis.

All cells in a table may not have observed counts. Sometimes, cells in a contingency table contain zero counts, known as empty cells. A zero count may be generated by two distinctive mechanisms. If sample size causes a zero count (e.g., Chinese American farmers in Illinois), it is known as a sampling zero. If, for a cell, it is theoretically impossible to have observations (e.g., male [contraceptive] diaphragm users), it is called a structural zero. Contingency tables with at least one structural zero are termed *incomplete tables*. When many cells in a contingency table have lower observed frequencies, be they zero or not, such a table is known as a SPARSE TABLE.

HISTORICAL DEVELOPMENT

Early work on CATEGORICAL DATA ANALYSIS in the beginning of the 20th century was primarily concerned with the analysis of contingency tables. The well-known debate between Karl Pearson and G. Udny Yule sparked interest in contingency table analysis. How would we analyze a 2×2 contingency table? Pearson was a firm believer in continuous bivariate distributions underlying the observed counts in the cross-classified table; Yule was of the opinion that variables containing discrete categories such as inoculation versus no inoculation would be best treated as discrete without assuming underlying distributions. Pearson's contingency coefficient was calculated based on the approximated underlying correlation of bivariate normal distributions collapsed into discrete cross-classifications, whereas Yule's Q was computed using a function of the ODDS RATIO of the observed discrete data directly.

Contributions throughout the 20th century can often be traced by the names of those inventing the statistics or tests for contingency table analysis. Pearson chi-square test, Yule's Q , Fisher's exact test, Kendall's tau, Cramér's V , Goodman and Kruskal's tau, and the Cochran-Mantel-Haenszel test (or the Mantel-Haenszel test) are just some examples. These statistics are commonly included in today's statistical software packages (e.g., proc freq in SAS generates most of the above).

Table 1 A Contingency Table of Parents' Socio-economic Status (SES) and Mental Health Status (expected values in parentheses)

Parent's SES	Mental Health Status			
	Well	Mild	Moderate	Impaired
		Symptom	Symptom	
A (high)	64 (48.5)	94 (95.0)	58 (57.1)	46 (61.4)
B	57 (45.3)	94 (88.8)	54 (53.4)	40 (57.4)
C	57 (53.1)	105 (104.1)	65 (62.6)	60 (67.3)
D	72 (71.0)	141 (139.3)	77 (83.7)	94 (90.0)
E	36 (49.0)	97 (96.1)	54 (57.8)	78 (62.1)
F (low)	21 (40.1)	71 (78.7)	54 (47.3)	71 (50.9)

Contingency table analysis can be viewed as the foundation for contemporary categorical data analysis. We may see the series of developments in LOG-LINEAR MODELING in the latter half of the 20th century as refinements in analyzing contingency tables. Similarly, the advancement in latent trait and latent class analysis is inseparable from the statistical foundation of contingency table analysis.

EXAMPLE

We present a classic data table, the midtown Manhattan data of mental health and parents' socio-economic status (SES), analyzed by L. Goodman, C. C. Clogg, and many others (see Table 1).

To analyze contingency tables, we need expected frequencies. Let F_{ij} indicate the expected value for the observed frequency f_{ij} in row i and column j in the table. The F_{ij} under the model of independence (no association between parents' SES and mental status) is given by

$$F_{ij} = \frac{f_{i+}f_{+j}}{f_{++}},$$

where f_{i+} gives the column total for the i th row, f_{+j} gives the row total for the j th column, and f_{++} gives the grand total of the entire table. For example, $f_{1+} = 64 + 94 + 58 + 46 = 262$, $f_{+1} = 64 + 57 + 57 + 36 + 21 = 307$, and $F_{11} = (262 \times 307)/1,660 = 48.5$. We compute the expected values for the entire table accordingly and present them in parentheses in Table 1.

The null hypothesis is that the two variables are unrelated or independent of each other. To test

the hypothesis, we compute the Pearson chi-square statistic and the LIKELIHOOD RATIO STATISTIC L^2 (or G^2):

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - F_{ij})^2}{F_{ij}} \quad \text{and}$$

$$L^2 = 2 \sum_i \sum_j f_{ij} \log \left(\frac{f_{ij}}{F_{ij}} \right).$$

Applying these formulas, we obtain a Pearson chi-square statistic of 45.99 and an L^2 of 47.42. With 15 degrees of freedom (= the number of rows minus 1 times the number of columns minus 1), we reject the null hypothesis of independence at any conventional significance level and conclude that mental health status and parents' SES are associated.

The model of independence can also be understood from the odds ratios computed from the expected frequencies. One may calculate the cross-product ratio for every adjacent 2×2 table contained in the contingency table to understand the independence assumption. That is, $(48.5 \times 88.8) / (95.0 \times 45.3) \approx 1.0$, $(95.0 \times 53.4) / (57.1 \times 88.8) \approx 1.0$, ..., $(57.8 \times 50.9) / (62.1 \times 47.3) \approx 1.0$. One quickly discovers that all such odds ratios are approximately 1! An ODDS RATIO of 1 indicates that there is no relationship between the two variables forming the table.

Both variables are ordinal in nature, but that information is not considered in the current analysis. Further fine-tuning via log-linear modeling will be necessary to analyze the type of association between the two variables.

COLLAPSIBILITY OF CONTINGENCY TABLES

A prominent issue in contingency table analysis is that of collapsibility. Simply put, a contingency table is collapsible if the fit of a particular statistical model is not significantly affected by combining its dimensions (i.e., removing variables), combining some categories in a variable (i.e., reducing the number of response categories), or both. In the current example, someone may suspect that there exists little difference between the mild and the moderate symptoms, noticing that the differences between the observed and the expected counts in the two middle columns in Table 1 are uniformly below 1. Table 2 presents the observed data (and the expected counts) with those two categories collapsed.

Table 2 A Contingency Table of Parents' Socio-economic Status (SES) and Mental Health Status With Collapsed Middle Symptom Categories (expected values in parentheses)

Parent's SES	Mental Health Status		
	Well	Mild or Moderate Symptom	Impaired
A (high)	64 (48.5)	152 (95.0)	46 (61.4)
B	57 (45.3)	148 (88.8)	40 (57.4)
C	57 (53.1)	170 (104.1)	60 (67.3)
D	72 (71.0)	218 (139.3)	94 (90.0)
E	36 (49.0)	151 (96.1)	78 (62.1)
F (low)	21 (40.1)	125 (78.7)	71 (50.9)

From Table 2, we arrive at a new set of test statistics: a Pearson chi-square of 43.51 and an L^2 of 44.94, with 10 degrees of freedom based on six rows and three columns (a still significant result at the .05 level). A likelihood ratio test based on either the Pearson chi-square statistics or the L^2 s by taking the difference of the new and the old values gives a test statistic of 2.47 with 5 (= 15 - 10) degrees of freedom. This result, which follows a chi-square distribution, is highly insignificant, suggesting that the two middle symptom categories can be collapsed without losing much information. The reader is encouraged to find out whether the parents' SES categories can be collapsed.

—Tim Futing Liao

REFERENCES

- Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley.
- Fienberg, S. E. (1981). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge: MIT Press.
- Rudas, T. (1997). *Odds ratios in the analysis of contingency tables*. Thousand Oaks, CA: Sage.

CONTINGENT EFFECTS

In theorizing about the relationship between an INDEPENDENT VARIABLE and a DEPENDENT VARIABLE, it may be the case that this relationship is contingent on the value of another independent variable. This type of relationship implies a multiplicative term in any

statistical model designed to empirically estimate the relationships of interest. Formally, this would take the following form:

$$Y = B_{12}(X_1 \cdot X_2),$$

where Y is the dependent variable, X_1 is an independent variable, X_2 is the variable on which the relationship between X_1 and Y is contingent, and B_{12} is the parameter that defines this relationship. Thus, the relationship between X_1 and Y is contingent on the value of X_2 . An example of this comes from Douglas Hibbs's (1982) classic work on presidential popularity in the United States during the Vietnam War era. Hibbs theorized that individuals' opinions toward the job that the president of the United States was doing would be directly affected by the number of U.S. military casualties that had occurred in recent weeks. In addition, he theorized that the magnitude of this effect would be contingent on individuals' social class because U.S. military casualties during the Vietnam era were disproportionately concentrated among the lower social classes.

A secondary definition for contingent effects comes from the standard inferences of multivariate models. For instance, consider the following nonstochastic component of a multivariate ORDINARY LEAST SQUARES regression with two independent variables:

$$Y = B_0 + B_1X_1 + B_2X_2,$$

where Y is the dependent variable, X_1 and X_2 are independent variables, and B_0 is the intercept term to be estimated along with the other parameters of interest. B_1 is the parameter (to be estimated) of the effect of X_1 on Y contingent on the effect of X_2 on Y . B_2 is the parameter (to be estimated) of the effect of X_2 on Y contingent on the effect of X_1 on Y . Another way that this is often expressed is that each slope parameter (B_1 and B_2 are the slope parameters in this example model) is the effect of one independent variable on the dependent variable, "controlling for" all other independent variables in the model.

—Guy Whitten

See also INTERACTION EFFECT, MULTIPLICATIVE

REFERENCE

Hibbs, D. A. (1982). The dynamics of political support for American presidents among occupational and partisan groups. *American Journal of Political Science*, 26, 312–332.

CONTINUOUS VARIABLE

A continuous variable is one within whose range of values any value is possible. Sometimes, it is also known as a metric variable or a quantitative variable. In practice, integer variables are often considered continuous.

—Tim Futing Liao

See also LEVEL OF MEASUREMENT

CONTRAST CODING

Contrast coding is a technique for coding $j - 1$ DUMMY VARIABLES to fully represent the information contained in a j -category NOMINAL classification. Contrast-coded dummy variables allow the researcher to aggregate categories into meaningful groupings for comparison purposes and, at the same time, test to determine whether component categories of each group are significantly different from each other. Coefficients for contrast-coded dummy variables provide comparisons of subgroup means sequentially disaggregated from *theoretically interesting* group combinations.

To illustrate, consider the relationship between marital status (married, separated, divorced, widowed, never married) and hourly wages (in U.S. dollars). In this example, marital status consists of five (J) subgroups, represented by four contrasts ($J - 1$). Although several initial aggregations are reasonable, we construct the first contrast as one between people who are currently legally married (married and separated) and those who are not (never married, divorced, and widowed), as illustrated in Table 1. Each successive contrast will disaggregate these two groups into component subgroups. Contrast 2 compares those separated to the married respondents. Because the "not married" aggregate contains three subgroups, we choose to compare the never-marrieds to a combination of respondents who are divorced or widowed in Contrast 3, with those divorced compared to the widowed in Contrast 4.

Two additional rules apply when using contrast codes for dummy variables. First, the codes within any single contrast-coded variable must sum to zero.

Table 1

Groups of Interest	New Variables Status				Means (U.S.\$)
	Contrast 1	Contrast 2	Contrast 3	Contrast 4	
Married	-1/2	-1	0	0	15.33
Never married	1/3	0	-1	0	12.95
Separated	-1/2	1	0	0	11.64
Divorced	1/3	0	1/2	1	12.94
Widowed	1/3	0	1/2	-1	11.93
Sum	-1/2 + 1/3 - 1/2 + 1/3 + 1/3 = 0	-1 + 0 + 1 + 0 + 0 = 0	0 - 1 + 0 + 1/2 + 1/2 = 0	0 + 0 + 0 + 1 - 1 = 0	Grand mean 14.26
Independence	C1 and C2 = $(-1/2)(-1) + 1/3(0) + (-1/2)(1) + (1/3)(0) + (1/3)(0) = 0$ C1 and C3 = $(-1/2)(0) + (1/3)(-1) + (-1/2)(0) + (1/3)(1/2) + (1/3)(1/2) = 0$ C1 and C4 = $(-1/2)(0) + (1/3)(0) + (-1/2)(0) + (1/3)(1) + (1/3)(-1) = 0$ C2 and C3 = $(-1)(0) + (0)(-1) + (1)(0) + (0)(1/2) + (0)(1/2) = 0$ C2 and C4 = $(-1)(0) + (0)(0) + (1)(0) + (0)(1) + (0)(-1) = 0$ C3 and C4 = $(0)(0) + (-1)(0) + (0)(0) + (1/2)(1) + (1/2)(-1) = 0$				

Second, the codes across dummy variables must be orthogonal. To test independence, we sum the products of successive pairs of codes, as illustrated in the bottom panel of the table, titled "Independence."

In general, descriptive statistics and zero-order correlations involving these dummy variables are not useful (Hardy, 1993). When contrast-coded dummy variables are used as independent variables in ORDINARY LEAST SQUARES REGRESSION, the constant reports the unweighted mean of the subgroup means, calculated as $[\sum_{j=1}^J \bar{X}_j] \div J$, the same as in regression with effects-coded dummy variables but unlike regression with BINARY coded dummy variables. The REGRESSION COEFFICIENTS can be used to calculate the difference in means within each contrast using the following formula: $C_j = B((n_{g1} + n_{g2})/(n_{g1})(n_{g2}))$, in which n_{g1} is the number of subgroups included in the first group, n_{g2} is the number of subgroups included in the second group, and B is the regression coefficient for the dummy variable of interest.

Results from a regression of hourly wage on the four contrast-coded dummy variables plus a constant are reported in Table 2. The value of the constant is 12.96, which one can verify as the unweighted mean of subgroup means by summing the category means provided in the previous table and dividing by 5. To evaluate the regression coefficients, consider the coefficient for Contrast 1. Substituting into the above equation, we have the following: Contrast 1 = $-1.054((2 + 3)/(2 \cdot 3)) = -.878$. This value is the difference between the unweighted mean of

Table 2

Model	Unstandardized Coefficients			
	B	Standard Error	t	Significance
Constant	12.96	.471	27.527	.000
Contrast 1	-1.054	.993	-1.061	.289
Contrast 2	-1.844	.353	-5.226	.000
Contrast 3	-.346	.772	-.448	.654
Contrast 4	.502	1.111	.452	.651

the subgroups of legally married couples $([15.33 + 11.64]/2 = 13.49)$ and the unweighted mean of the subgroups of people who are not legally married $([12.95 + 12.94 + 11.93]/3 = 12.61)$. People who are not married legally make, on average, \$.88 less than people who are legally married. However, the t -test for this coefficient is not statistically significant at conventional levels; therefore, the difference of \$.88 is not reliable. The appropriate inferential test to determine whether hourly wages differ by marital status is the F -test, not the t -TEST, because including information about the different marital statuses requires all four dummy variables.

—Melissa A. Hardy and Chardie L. Baird

See also DUMMY VARIABLE, EFFECTS CODING

REFERENCE

- Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage.

CONTROL

Control is concerned with a number of functions by which the existence of a relationship between two main variables, an independent (or explanatory) variable and a dependent (or response) variable, can be ascertained without the interference of other factors. The four main functions of control are to fix, or “hold constant,” CONFOUNDING effects; to test for spurious effects between the variables; to serve as a reference point in experimental tests; and to ascertain the *net* effect of the independent variable when other variables also have an effect.

Confounding effects could be ones that intervene in the bivariate relationship or ones that precede it. Alternatively, the relationship could be a spurious one, that is, when a third variable affects both the independent and dependent variables. To check for the effect of such confounders, we control for them, so that we are comparing like with like (see *CETERIS PARIBUS*). For example, we may notice that certain minority ethnic groups in Britain achieve less well at school than White children. We might, therefore, posit that cultural or genetic factors associated with ethnicity cause some groups to fare worse than others at school. However, if we consider that there is also known to be an impact of income and deprivation on school achievement levels and that certain minority groups tend to be concentrated at the bottom of the income bracket, we might wish to control for income levels. We would then find that the relationship is not sustained and that, controlling for income (i.e., comparing poor kids with poor kids and rich kids with rich kids), children from Bangladeshi backgrounds fare at least as well as their White counterparts. The apparent relationship between ethnicity itself and low achievement is thus not sustained when background is controlled for. This example of an intervening effect is illustrated in Figure 1. Bangladeshi children still have poor educational outcomes, but the process of causal reasoning has been modified, placing the onus of explanation on their poverty rather than their ethnicity.

An alternative example of an intervening relationship could deal with women’s lower average earnings rather than men’s. An initial relationship might be posited between sex and lower earnings, which focuses on discrimination within the workplace. It is relevant to consider that women tend to be concentrated in the lower echelons of the employment hierarchy. If

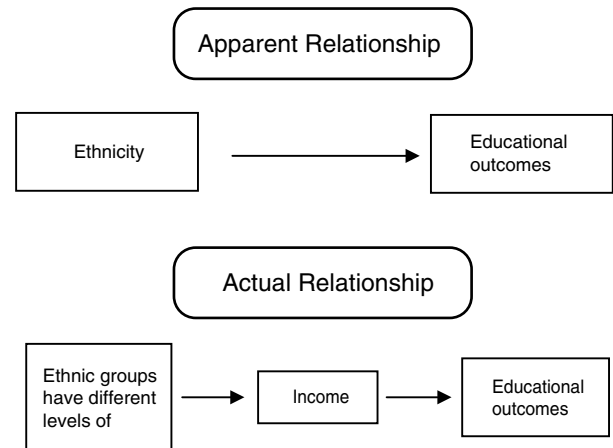


Figure 1 The Effect of Controlling for Income in Considering the Relationship Between Ethnicity and Educational Outcomes

we hold occupational class constant across men and women, do we still see a gender effect? What we find is that the effect remains, but it is much reduced. Women’s occupational structure explains in part their lower earnings rather than their sex per se, and this is revealed by comparing men and women of the same occupational class. See Figure 2 for an illustration of this relationship.

Additional relevant factors to those of interest can be controlled for through experimental design or by using statistical methods. In experimental design, the effect of an intervention can be tested by randomly applying the intervention to one subset of test cases and withholding it from the remaining CONTROL GROUP. The control group can also be selected to match the “experimental group” in critical ways (such as age, sex, etc.). For example, medical interventions are typically tested using RANDOMIZED CONTROLLED TRIALS. *Random* refers to the fact that those who are experimented on are randomly selected from within the population of interest, and *controlled* means that for every case that is selected for EXPERIMENT, a parallel case will not receive the intervention but will be treated in every other way as the experimental group, including, possibly, being given a placebo treatment. Such trials are double blind if not only the experiment and control groups do not know which group they belong to but also even the practitioner administering the intervention (or placebo) is unaware, thus minimizing the impact of researcher bias. In such experiments, given random allocation of the intervention, any differences

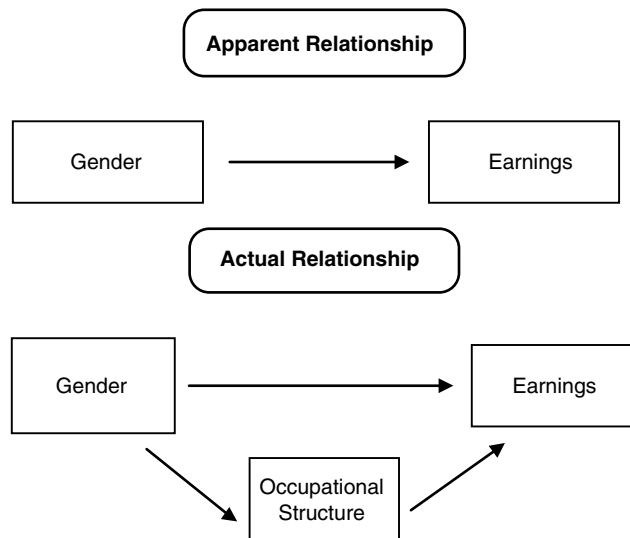


Figure 2 The Effect of Holding Occupational Class Constant When Assessing the Relationship Between Gender and Earnings

in outcomes between the experimental group and the control group can then be attributed to the intervention.

Although experimental techniques are regularly used in psychology, more commonly in the social sciences, it is not practical and would not be meaningful to run experiments. For example, it would not be feasible to establish an experiment on the effects of parental separation on child outcomes, which would require a random sample of the experimental group to separate from their partners and the matched control group to remain with them. Instead, confounding effects are controlled for using statistical methods. There are a variety of multivariate techniques for assessing the effect of a number of different variables simultaneously and thus being able to distinguish the net effect of one when another (or several others) is controlled for. Regression analysis, such as MULTIPLE REGRESSION ANALYSIS or LOGISTIC REGRESSION, is the most commonly used way of ascertaining such net effects. In such analysis, it may be that additional variables are included in a model that do not show any significant effect themselves, simply to ensure that all relevant factors have been controlled for.

—Lucinda Platt

See also CETERIS PARIBUS, LOGISTIC REGRESSION, MULTIVARIATE ANALYSIS, MULTIPLE REGRESSION ANALYSIS

REFERENCE

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

CONTROL GROUP

A control group is the group with which an experimental group (or treatment group) is contrasted. It consists of units of study that either did not receive the treatment, whose effect is under investigation, or units of study that did receive a placebo, whose effect is not being studied. Control groups may be alternatively called baseline groups or contrast groups. They are formed through random assignment of units of study, as in between-subject designs, or from the units of study themselves, as in WITHIN-SUBJECT DESIGNS. In true-experimental studies and between-subject designs, units of study (such as children, clinical patients, schools, hospitals, communities, etc.) are first randomly selected from their respective populations; then they are randomly assigned into either a control group or an experimental group. At the conclusion of the study, outcome measures (such as scores on one or more dependent variables) are compared between those in the control group and those in the experimental group (or groups). The effect of a treatment is assessed on the basis of the difference (or differences) observed between the control group and one or more experimental group. In true-experimental studies and within-subject designs, units of study are also randomly selected from their respective populations, but they are not randomly assigned into control versus experimental groups. Instead, baseline data are gathered from units of study themselves. These data are treated as “control data,” to be compared with outcome measures that are hypothesized to be the result of a treatment. Thus, units of study act as their own “control group” in within-subject designs. For QUASI-EXPERIMENTAL studies, treatments are not administered by an experimenter to units of study or participants, as in true-experimental studies. Treatments are broadly construed to be characteristics of participants, such as gender, age, and socioeconomic status (SES), or features of their settings, such as public versus private school and single- or intact-family background. Quasi-experimental studies may also be necessary when researchers have little or no control over randomly assigning participants into

different treatment conditions. In this case, a nested design is appropriate and necessary.

—Chao-Ying Joanne Peng

See also EXPERIMENT

REFERENCES

- Huck, S. (2000). *Reading statistics and research* (3rd ed.). New York: Addison-Wesley Longman.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.
- Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

CONVENIENCE SAMPLE

Convenience sampling, as its name suggests, involves selecting sample units that are readily accessible to the researcher. It is also sometimes called accidental sampling and is a form of NONPROBABILITY SAMPLING; that is, each member of a population has an unknown and unequal probability of being selected.

The advantages of convenience samples are that they are relatively inexpensive and, by definition, easy to access. Sometimes, this form of sampling may be the most efficient way to access a hard-to-reach population, such as sex workers or illicit drug users. Convenience sampling becomes more appropriate and acceptable when the POPULATION of interest is difficult to define to allow any reliable form of random sampling. It is also deemed more legitimate when gaining access to the population of interest is difficult.

Although studies using convenience samples may yield intriguing findings, they suffer from the inability to generalize beyond the samples. Researchers will not know whether the sample is representative of the population under study. On the basis of only a convenience sample, researchers will not know whether their sample is typical or atypical of other groups. However, this shortfall is common to all nonprobability samples. Researchers using this form of sampling should also consider the potential BIAS involved.

The use of convenience sampling is not limited to any particular subject area. Researchers have used convenience samples to address issues of ageism, examine sexual decisions made by inner-city Black adolescents, and explore the stigma of wheelchair use and how the public reacts to wheelchair users.

Convenience sampling is also employed in both qualitative and quantitative data collection. A typical example of employing convenience sampling in quantitative data collection is surveying students in a particular course, generally a course taught by the researcher or a friend of the researcher. The students are essentially “captured” participants in a class setting, easily accessible by the researcher. A typical example of using convenience sampling in collecting qualitative information is when researchers conduct in-depth interviews with family members. Most family members, especially those living in the same household, are likely to participate in the interviews as they are likely to have a more trusting relationship with the interviewer than with a stranger. Most of all, they are also readily available for the interview.

—VoonChin Phua

CONVERSATION ANALYSIS

When people talk with one another, they are not merely communicating thoughts, information, or knowledge: In conversation, as in all forms of interaction, people are *doing* things in talk. They are engaged in social activities with one another, such as blaming, complaining, inviting, instructing, telling their troubles, and so on. Conversation analysis (CA) is the study of how participants in talk-in-interaction organize their verbal and nonverbal behavior so as to conduct such social actions and, more broadly, their social affairs. It is therefore primarily an approach to social action (Schegloff, 1996). Methodologically, it seeks to uncover the practices, patterns, and generally the methods through which participants perform and interpret social action.

CA originated with the work that Harvey Sacks (1935–1975) undertook at the Center for the Scientific Study of Suicide, in Los Angeles, 1963–1964. Drawn by his interests both in the ethnomethodological concern with members’ methods of practical reasoning (arising from his association with Harold Garfinkel) and in the study of interaction (stimulated by his having been taught as a graduate student at Berkeley by Erving Goffman), Sacks began to analyze telephone calls made to the Suicide Prevention Center (SPC). Without any diminished sensitivity to the plight of persons calling the SPC, Sacks investigated how callers’ accounts of their troubles were produced in

the course of their conversations over the telephone with SPC counselors. This led him to explore the more generic “machineries” of conversational turn taking, along with the sequential patterns or structures associated with the management of activities in conversation. Through the collection of a broader corpus of interactions, including group therapy sessions and mundane telephone conversations, and in collaboration with Gail Jefferson (1938–) and Emanuel Schegloff (1937–), Sacks began to lay out a quite comprehensive picture of the conversational organization of turn taking; overlapping talk; repair; topic initiation and closing; greetings, questions, invitations, requests, and so forth and their associated sequences (adjacency pairs); agreement and disagreement; storytelling; and integration of speech with nonvocal activities (Sacks, 1992; for a comprehensive review of CA’s topics and methodology, see ten Have, 1999). Subsequent research in CA over the past 30 years has shown how these and other technical aspects of talk-in-interaction are structured, socially organized resources—or methods—whereby participants perform and coordinate activities through talking together. Thus, they are the technical bedrock on which people build their social lives or, in other words, construct their sense of sociality with one another.

In many respects, CA lies at the intersection between sociology and other cognate disciplines, especially linguistics and social psychology. Certainly, research in CA has paralleled developments within sociolinguistics, pragmatics, discourse analysis, and so forth toward a naturalistic, observation-based science of actual verbal behavior, which uses recordings of naturally occurring interactions as the basic form of data (Heritage, 1984). Despite the connections between CA and other disciplines, CA is distinctively sociological in its approach in the following kinds of ways. First, in its focus on how participants understand and respond to one another in their turns at talk, CA explores the social and interactional underpinnings of intersubjectivity—the maintenance of common, shared, and even “collective” understandings between social actors. Second, CA develops empirically Goffman’s (1983) insight that social interaction embodies a distinct moral and institutional order that can be treated like other social institutions, such as the family, economy, religion, and so on. By the “interaction order,” Goffman meant the institutional order of interaction, and CA studies the practices that make up this institution as a topic in its own right. Third, all levels of linguistic production (including syntactic, prosodic, and phonetic) can be

related to the actions (such as greetings, invitations, requests) or activities (instructing, cross-examining, performing a medical examination and diagnosing, etc.) in which people are engaged when interacting with one another. In this way, conversational organizations underlie social action (Atkinson & Heritage, 1984); hence, CA offers a methodology, based on analyzing sequences in which actions are produced and embedded, for investigating how we accomplish social actions. This is applicable equally to both verbal and nonverbal conduct, as well as the integration between them, in face-to-face interaction (e.g., Goodwin, 1981). Fourth, it is evident that the performance by one participant of certain kinds of actions (e.g., a greeting, question, invitation, etc.) sets up certain expectations concerning what the other, the recipient, should do in response; that is, he or she may be expected to return a greeting, answer the question, accept or decline the invitation, and so on. Thus, such pairs of actions, called *adjacency pairs* in CA, are normative frameworks within which certain actions should properly or accountably be done: The normative character of action and the associated accountability of acting in accordance with normative expectations are vitally germane to sociology’s sense of the moral order of social life, including ascriptions of deviance. Fifth, CA relates talk to social context. CA’s approach to context is distinctive, partly because the most proximate context for any turn at talk is regarded as being the (action) sequence of which it is a part, particularly the immediately prior turn. Also, CA takes the position that the “context” of an interaction cannot be exhaustively defined by the analyst a priori; rather, participants display their sense of relevant context (mutual knowledge, including what each knows about the other; the setting; relevant biographical information; their relevant identities or relationship, etc.) in the particular ways in which they design their talk—that is, in the recipient design of their talk.

Finally, as a method of enquiry, CA is equally as applicable to institutional interactions as it is to mundane conversation (Drew & Heritage, 1992). The social worlds of the law, medicine, corporations and business negotiations, counseling, education, and public broadcast media (especially news interviews) and other such institutional and workplace settings are conducted through talk-in-interaction. CA has increasingly come to focus on how participants in such settings manage their respective institutional activities, for instance, as professionals (doctors, lawyers, etc.) or as laypersons

(clients, patients, students, etc.). Methodologically, the objective of CA research into institutional interactions is to reveal the practices through which participants manage their interactions as a specialized form of interaction—for instance, as a news interview (Clayman & Heritage, 2002), as doing counseling, as courtroom cross-examination, and so forth. In such research, CA connects very directly with ethnographic sociology and offers a distinctive, rigorous, and fruitful methodology for investigating the organizational and institutional sites of social life.

—Paul Drew

REFERENCES

- Atkinson, J. M., & Heritage, J. (Eds.). (1984). *Structures of social action: Studies in conversation analysis*. Cambridge, UK: Cambridge University Press.
- Clayman, S., & Heritage, J. (2002). *The news interview: Journalists and public figures on the air*. Cambridge, UK: Cambridge University Press.
- Drew, P., & Heritage, J. (Eds.). (1992). *Talk at work: Interaction in institutional settings*. Cambridge, UK: Cambridge University Press.
- Goffman, E. (1983). The interaction order. *American Sociological Review*, 48, 1–17.
- Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.
- Heritage, J. (1984). *Garfinkel and ethnomethodology*. Cambridge, UK: Polity.
- Sacks, H. (1992). *Lectures on conversation* (2 vols.). Oxford, UK: Blackwell.
- Schegloff, E. A. (1996). Confirming allusions: Toward an empirical account of action. *American Journal of Sociology*, 2, 161–216.
- ten Have, P. (1999). *Doing conversation analysis*. Thousand Oaks, CA: Sage.

CORRELATION

Generally speaking, the notion of correlation is very close to that of ASSOCIATION. It is the statistical association between two variables of interval or ratio measurement level.

To begin with, correlation should not be confused with causal effect. Indeed, statistical research into causal effect for only two variables happens to be impossible, at least in observational research. Even in extreme cases of so-called unicausal effects, such as contamination with the *Treponema pallidum*

bacteria and contracting syphilis, there are always other variables that come into play, contributing, modifying, or counteracting the effect. The presence of a causal effect can only be sorted out in a multivariate context and is consequently more complex than correlation analysis.

Let us limit ourselves to two interval variables, denoted as X and Y , and let us leave CAUSALITY aside. We assume for the moment a linear model. It is important to note that there is a difference between the correlation coefficient, often indicated as r , and the UNSTANDARDIZED regression coefficient b (computer output: B). The latter merely indicates the slope of the REGRESSION line and is computed as the tangent of the angle formed by the regression line and the x -axis. With income (X) and consumption (Y) as variables, we now have the consumption quote, which is the additional amount one spends after obtaining one unit of extra income, so the change in Y per additional unit in X is $B = \Delta Y / \Delta X$. We will see below that there are, in fact, two such regression coefficients and that the correlation coefficient is the geometrical MEAN of them.

FIVE MAIN FEATURES OF A CORRELATION

Starting from probabilistic correlations, each correlation has five main features:

1. nature,
2. direction,
3. sign,
4. strength,
5. statistical generalization capacity.

1. Nature of the Correlation

The nature of the correlation is linear for the simple correlation computation suggested above. This means that through the scatterplot, a linear function of the form $E(Y) = b_0 + b_{y1}X_1$ is estimated. Behind the correlation coefficient of, for example, $r = 0.40$ is a linear model. Many researchers are fixated on the number between 0 and 1 or between 0 and -1 , and they tend to forget this background model. Often, they unconsciously use the linear model as a tacit obviousness. They do not seem to realize that nonlinearity occurs frequently.

An example of nonlinearity is the correlation between the percentage of Catholics and the percentage of CDU/CSU voters in former West Germany (Christlich-Demokratische Union/Christlich-Soziale

Union). One might expect a linear correlation: the more Catholics, the more voters of CDU/CSU according to a fixed pattern. However, this “the more, the more” pattern only seems to be valid for communes with many Catholics. For communes with few Catholics, the correlation turns out to be fairly negative: The more Catholics, the fewer voters for CDU/CSU. Consequently, the overall scatterplot displays a U pattern. At first, it drops, and from a certain percentage of Catholics onwards, it starts to rise. The quadratic function that describes a parabola therefore shows a better fit with the scatterplot than the linear function.

Many other nonlinear functions describe reality, whether exponential, logistic, discontinuous, or other. The exponential function, for example, was used in the reports by the Club of Rome, in which increased environmental pollution (correlated with time) was described in the 1970s. The logistic function was used by those who reacted against the Club of Rome, with the objection that the unchecked pollution would reach a saturation level.

2. The Direction of the Correlation

When dealing with two variables, X_1 and Y , one of them is the INDEPENDENT VARIABLE and the other is the DEPENDENT VARIABLE. In the case of income (X_1) and consumption (Y) the direction $X_1 \rightarrow Y$ is obvious. Here, a distinction between the correlation coefficient and the regression coefficient can already be elucidated. The regression coefficient indicates b_{y1} for the direction $X_1 \rightarrow Y$ and b_{1y} for the direction $Y \rightarrow X_1$. On the other hand, the correlation coefficient, which can be calculated as the geometrical mean of b_{y1} and b_{1y} (= the square root of the product of these coefficients) incorporates both directions and therefore is a nondirected symmetrical coefficient.

A novice might conclude that causal analysis is the privilege of regression coefficients and that correlation coefficients are unsuitable for that purpose. This would be an understandable but premature conclusion. Indeed, the direction in which we calculate mathematically is not the same as the causality direction. Each function $Y = aX + b$, with Y as a dependent variable, is convertible into its inverted function $X = Y/a - b/a$, with X as a dependent variable. The direction here is of relative value. However, when a stone is thrown into a pond, the expanding ripples are explained by the falling stone, but the opposite process with shrinking ripples would imply an enormous number of ripple generators,

the coherence of which would need to be explained. It would also be imperative to demonstrate that they come from one central point (the improbability of an implosion). The direction here is not of relative value.

As was said before, causality cannot be equated to the statistical computation of a regression coefficient. True, it is fortunate that this coefficient represents a direction, and Hubert Blalock (1972) gratefully used this in his CAUSAL MODEL approach. However, this direction in terms of causality is hopelessly relative in statistical computation. This is admitted with perfect honesty by Sewell Wright (1934) in his article “The Method of Path Coefficients” and in earlier articles, when he claimed to have found a method to link the (a priori) causal knowledge with the (a posteriori) empiric correlations in the correlation matrix.

3. The Sign of the Correlation

A correlation can be positive (the more of X_1 , the more of Y) or negative (the more of X_1 , the less of Y). In the case of income and consumption, there is a positive correlation (the sign is +): The higher the income, the higher the consumption.

An example of negative correlation is the relationship between social class and number of children at the beginning of the past century. Higher social classes displayed the mechanism that was known as *social capillarity*. This was strikingly described by Friedrich Engels in *Der Ursprung der Familie, des Privateigentums und des Staates*. As soon as a son was born in the higher social classes, the parents started applying birth control because the estate could be passed on from father to son. In the lower social classes, however, this mechanism was lacking. A family of 10 children was the rule, and sometimes there were as many as 20 children. The sign here is abundantly clear—the higher the social class, the lower the number of children.

4. The Strength of the Correlation

A correlation can be weak or strong. A correlation coefficient of 0.90 or -0.90 is strong. A coefficient of 0.10 or -0.10 is weak. A coefficient 0 signals lack of correlation. A coefficient 1 or -1 indicates maximal deterministic correlation.

It is convenient to keep the vector model in mind. The vector model is the coordinate system in which the roles of variables and units have been changed:

Units are now the axes, and variables are the points forming the scatterplot. These variables are conceived as the end points of vectors starting in the origin—hence the name. In such a vector model, the correlation coefficient between two variables is equal to the cosine of the angle formed by the two vectors. We know that the cosine of 90° equals 0: The vectors then are at perpendicular angles to each other, and the correlation is 0. On the other hand, the cosine of 0° equals 1: The vectors then coincide and are optimally correlated.

Mostly, a strong correlation coefficient (between two variables, not in a multivariate context) will coincide with a strong regression coefficient, but this is not always the case. Many situations are possible, and this is related to the dispersions of the variables, which will be dealt with below.

In addition, a strong correlation will tend to be more significant than a weak correlation, but it does happen that a strong correlation is not significant. Moreover, a significant correlation can be weak. It follows that the strength of a correlation (correlation or regression coefficient) cannot be judged as such because it should be studied in combination with other criteria, the first one being the significance, which is discussed next.

5. The Statistical Potential for Generalization

A correlation calculated on the basis of a random sample may or may not be significant (i.e., it can or cannot be statistically generalized with respect to the population from which the sample was taken). Here we have to go against the postmodern range of ideas and take a traditional rigorous point of view: Only those correlations that can be statistically generalized (and are therefore statistically significant) are appropriate for interpretation. Nonsignificant correlations are the result of mere randomness and should therefore be considered nonexistent for the population.

Consequently, a strong coefficient (e.g., $r = 0.90$), which is not significant because the sample size is much too small (e.g., $n = 5$), does not stand the test and will be rejected by any social science researcher. A discerning mind will object that a weak correlation (e.g., $r = 0.10$) in a sufficiently large sample (e.g., sample size $n = 5,000$) is still significant and that this is only due to the size of the sample. Take a sufficiently large sample and you can prove anything with statistics, so it would appear. However, this argument is

only partly correct. When dealing with large numbers, weak correlations or small changes such as changes in fertility can be extremely significant (in the sense of important) for policymaking. Any demographer knows how irreversible such processes are. Besides, for large numbers, it is also still possible that the correlation is not statistically significant. Therefore, it does not hold to pretend that the result is only due to the size of the sample.

Admittedly, there is still a grey area between two extremes: a significant result for a small sample and a nonsignificant one for a large sample. In the latter case, we know with certainty that the correlation is not significant. In the former case, we are certain that a larger sample survey will also yield a significant result. However, in general and certainly in the grey area, non-statistical criteria, such as social relevance, will always contribute to the interpretation of correlation results.

So far, we have looked at some properties of correlation: nature, direction, sign, strength, and significance. As was already mentioned, one should not be blinded here by the number representing the correlation coefficient; indeed, all these features need to be taken into account. In this sense, major META-ANALYSES in which correlation coefficients from different studies are all mixed together should not be trusted. Indeed, in each separate study, one should raise many questions. Are we dealing with a linear correlation? If so, is the regression coefficient different from the correlation coefficient, and how large is the intercept? Are the variables standardized? Is the correlation significant? How large is the confidence interval? What is the probability of TYPE I ERROR α ? What is the size of the sample? What is the probability of TYPE II ERROR β and the power?

THE DISPERSION OF THE VARIABLES: STANDARDIZED COEFFICIENTS?

The question about the difference between correlation and regression effect actually also refers to the DISPERSION of the variables. Indeed, we have already mentioned that a strong correlation coefficient does not always coincide with a strong regression coefficient and that this is linked to the dispersion of the variables. This can be derived from the formulas of the coefficients: r_{y1} (or r_{1y}) is the covariance divided by the geometrical mean of the variances, b_{y1} is the covariance divided by the variance of X_1 , and b_{1y} is the covariance divided by the variance of Y . In the

case of equal dispersions, all three coefficients will be equal. A greater dispersion of Y will have a decreasing effect on the correlation coefficient and on b_{1y} , but not on b_{y1} .

With respect to this, one may wonder if the standardized coefficients (betas) should be preferred. (Standardization is a procedure of subtracting the mean and dividing by the STANDARD DEVIATION.) Dividing by the standard deviation happens to be a rather drastic operation, which causes the variance of the variables to be artificially set equal to 1, with a view to obtaining comparability. But some have argued that a similar difference in education has a different meaning in a society with a great deal of educational inequality compared to a society with a small degree of such inequality, and therefore they advocate nonstandardized coefficients.

All this is not just abstract thinking; it does occur in specific processes in society. Examples may include the processes of efficient birth control, underemployment, the raising of political public opinion, and environmental awareness. From all these examples, we learn that society introduces changes in the dispersion of its features.

CALCULATION OF THE CORRELATION COEFFICIENT

We would like to illustrate this importance of dispersions with a small example in which the correlation coefficient between education (X) and income (Y) is calculated. The formula is $r_{xy} = [\sum(X - \bar{X})(Y - \bar{Y})] / [\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2]^{1/2}$.

Let us take a nation in which few of its citizens go to school, and those who do happen to be educated have high incomes and the rest are poor. Let the data matrix be as below. Education is measured on a scale from 1 to 9, and income is measured in thousands of Euros.

X	1	1	1	2	2	3	3	3	4	4	6	6	6	7	7	8	8	8	9	9
Y	2	4	5	3	5	3	4	6	4	5	5	7	8	6	8	6	7	9	7	8

The scatterplot is shown in Figure 1.

The standard deviations are 2.81 for education and 1.93 for income.

The correlation coefficient is 0.798. The model is linear, the coefficient is symmetric, the relationship is positive, the strength is substantial, and the coefficient is statistically significant ($p < .001$).

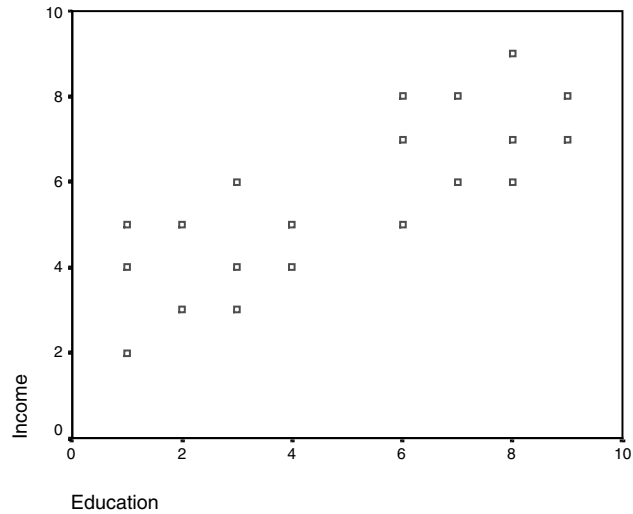


Figure 1

Now suppose that in this nation, the more ambitious among the poor demand that their children be admitted to the schools and that the state yields to this pressure and begins to subsidize education. Large numbers now go to school and college, average income rises as education raises the general level of human capability, and the variation of incomes declines. But once subsidized education becomes widespread, the individuals who are more schooled can no longer capture for themselves much of the benefit of that schooling, and the correlation between income and education diminishes.

Now the data matrix is the following.

X	6	6	6	6	6	7	7	7	7	8	8	8	8	8	8	9	9	9	9	
Y	5	7	8	5	7	8	6	8	6	8	6	7	9	6	7	9	7	8	7	8

The scatterplot is shown in Figure 2.

The standard deviations are much smaller: 1.14 for education and 1.17 for income. The correlation coefficient has diminished from 0.798 to 0.285. The relationship is positive but small and no longer statistically significant.

One could also suppose that in this nation, the variation of education remains the same because only some of the citizens go to school, but the variation of incomes has declined because there is a lot of underemployment (people with high diplomas having low-salary jobs). This would result in the following data matrix:

X	1	1	1	2	2	3	3	3	4	4	6	6	6	7	7	8	8	8	9	9
Y	2	4	5	3	5	3	4	6	4	5	2	4	5	3	5	3	4	6	4	5

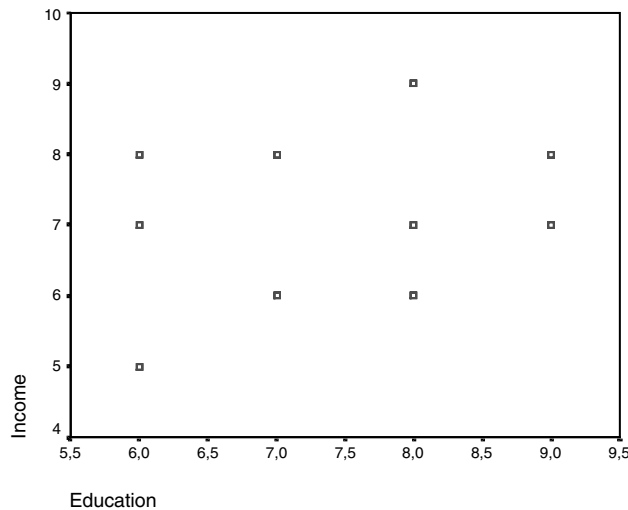


Figure 2

The scatterplot is now as follows, in Figure 3:

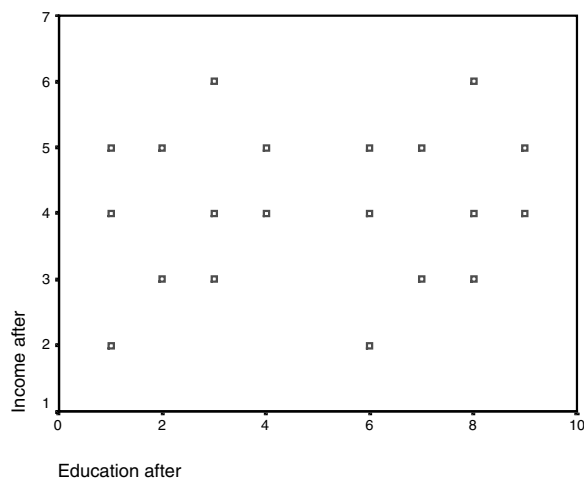


Figure 3

The standard deviation of education remains the same (2.81), but the standard deviation of income has become very small (1.17). The correlation coefficient is now 0.116. The relationship is positive but small and nonsignificant.

—Jacques Tacq

REFERENCES

Blalock, H. M. (1972). *Social statistics*. Tokyo: McGraw-Hill Kogakusha.

- Brownlee, K. (1965). *Statistical theory and methodology in science and engineering*. New York: John Wiley.
- Hays, W. L. (1972). *Statistics for the social sciences*. New York: Holt.
- Kendall, M., & Stuart, A. (1969). *The advanced theory of statistics*. London: Griffin.
- Tacq, J. J. A. (1997). *Multivariate analysis techniques in social science research: From problem to analysis*. London: Sage.

CORRESPONDENCE ANALYSIS

Correspondence analysis is a way of *seeing* the ASSOCIATION in a two-way cross-tabulation rather than *measuring* it. Consider, for example, the cross-tabulation in Table 1, taken from the 1994 International Social Survey Programme (ISSP) on Family and Changing Gender Roles. This table shows, for 24 different countries, how many respondents think that the ideal number of children in a family is 0, 1, 2, 3, 4, or 5 or more. Response percentages for each country are given in parentheses: For example, in Australia, 833 of the 1,529 respondents, or 54.5%, considered two children to be ideal.

The usual Pearson CHI-SQUARE statistic for this 24×6 table has the value 6,734, and CRAMÉR'S V is 0.206, both of which are highly significant. In fact, any MEASURE OF ASSOCIATION, even with a low value, will almost always be highly significant for a table with such high frequencies. Clearly, there are significant differences between countries, but what are these differences? We can see from the table that the United States and Canada are quite similar, but how can we compare their positions on this issue with the European countries? Correspondence analysis can answer this question because it aims to show the similarities and differences between the countries in a compact, graphical way.

The correspondence analysis of Table 1 is shown in Figure 1. This is a two-dimensional "map" of the positions of the row and column points, showing the main features of the contingency table. The points representing the columns form an arch pattern, starting from the bottom right and ending at the bottom left. This indicates a monotonic ordination of the countries in terms of their opinions on this issue, from those wanting fewer children (e.g., [former] East Germany, Bulgaria, and the Czech Republic) to those wanting more children (e.g., Ireland, Israel, and the Philippines). The horizontal axis, or first principal axis, captures this

Table 1 Ideal Number of Children in Family: Response Frequencies in 24 Countries (percentages in parentheses)

Country	0	1	2	3	4	≥ 5	Total
Australia	4 (0.3)	19 (1.2)	833 (54.5)	448 (29.3)	203 (13.3)	22 (1.4)	1,529
West Germany	18 (0.8)	116 (5.2)	1551 (69.5)	441 (19.8)	79 (3.6)	26 (1.2)	2,231
East Germany	6 (0.6)	108 (10.1)	831 (77.7)	115 (10.7)	10 (3.6)	0 (0.0)	1,070
Great Britain	5 (0.6)	16 (1.8)	679 (74.9)	145 (16.0)	55 (6.1)	6 (0.7)	906
Northern Ireland	0 (0.0)	7 (1.2)	271 (45.5)	157 (26.4)	142 (23.9)	18 (3.0)	595
United States	11 (0.8)	34 (2.6)	783 (60.0)	295 (22.6)	159 (12.2)	22 (1.7)	1,304
Austria	5 (0.5)	51 (5.4)	667 (70.4)	195 (20.6)	28 (3.0)	2 (0.2)	948
Hungary	8 (0.6)	70 (4.8)	839 (57.7)	485 (33.4)	41 (2.8)	11 (0.8)	1,454
Italy	4 (0.4)	46 (4.5)	678 (67.1)	254 (25.1)	22 (2.2)	7 (0.7)	1,011
Ireland	2 (0.2)	7 (0.8)	276 (31.3)	278 (31.5)	270 (30.6)	49 (5.6)	882
Netherlands	41 (2.1)	42 (2.1)	1050 (53.4)	575 (29.2)	199 (10.1)	61 (3.1)	1,968
Norway	3 (0.2)	10 (0.5)	939 (48.2)	824 (42.3)	145 (7.4)	29 (1.5)	1,950
Sweden	0 (0.0)	8 (0.7)	741 (63.9)	329 (28.4)	65 (5.6)	16 (1.4)	1,159
Czech Republic	7 (0.7)	109 (10.7)	688 (67.3)	192 (18.8)	16 (1.6)	11 (1.1)	1,023
Slovenia	9 (0.9)	40 (3.9)	601 (59.2)	311 (30.6)	45 (4.4)	9 (0.9)	1,015
Poland	0 (0.0)	27 (1.9)	756 (54.1)	483 (34.6)	99 (7.1)	32 (2.3)	1,397
Bulgaria	55 (4.9)	80 (7.1)	742 (65.9)	216 (19.2)	29 (2.6)	4 (0.4)	1,126
Russia	0 (0.0)	188 (9.6)	1,219 (62.5)	463 (23.7)	50 (2.6)	30 (1.5)	1,950
New Zealand	8 (0.9)	11 (1.2)	497 (55.7)	258 (28.9)	106 (11.9)	12 (2.3)	892
Canada	5 (0.4)	25 (2.2)	674 (59.3)	308 (27.1)	107 (9.4)	17 (1.5)	1,136
Philippines	0 (0.0)	10 (0.8)	229 (19.1)	561 (46.8)	282 (23.5)	117 (9.8)	1,199
Israel	1 (0.1)	9 (0.7)	204 (16.7)	525 (42.9)	384 (31.4)	100 (8.2)	1,223
Japan	0 (0.0)	9 (0.7)	483 (37.2)	737 (56.7)	56 (4.3)	15 (1.2)	1,300
Spain	0 (0.0)	140 (6.0)	1,441 (61.7)	593 (25.4)	108 (4.6)	55 (2.4)	2,337
Total	192 (0.6)	1,182 (3.7)	17,672 (55.9)	9,188 (29.1)	2,700 (8.5)	671 (2.1)	31,605

NOTE: The ISSP data on which Table 1 is based were made available by the *Zentralarchiv für Empirische Sozialforschung* in Cologne (www.gesis.org/en/data_service/issp/index.htm).

general scaling of the countries, whereas the vertical second principal axis contrasts the extreme categories with the middle categories. Such an arch pattern is frequently encountered in correspondence analysis maps. Most of the highly developed countries are around the center of the map, which represents the average, but there are interesting differences among them. The United States, for example, lies lower down on the second axis, mostly because of its relatively high frequencies of “4 children” compared to other countries around the center. Japan, by contrast, lies high up on the second axis because of its very high frequency of “3 children.”

Each country finds its position in the map according to its *profile*, or set of relative frequencies, shown as percentages in Table 1. The center of the map represents the *average profile*—that is, 0.6%, 3.7%, 55.9%, 29.1%, 8.5%, and 2.1% for all the countries (see last row of Table 1)—and a particular country will be attracted in the direction of a category if its percentage is higher than the corresponding average value.

The method can be equivalently defined as a variant of METRIC VARIABLE multidimensional scaling, in which distances between countries are measured by the so-called chi-square distances (Benzécri, 1973), and each country is weighted proportional to its respective sample size.

The main issue in constructing the map is to decide on which axis scaling to choose from several different options. Figure 1 is called a *symmetric map* because it represents the rows and columns in an identical way using their respective *principal coordinates*, so that each set of points has the same dispersion in the map. An *asymmetric map* is also possible when each country is located at the weighted average position of the response category points, with the latter being represented in *standard coordinates* (Greenacre, 1993). The VARIANCE in the data table is measured by the *inertia*, equal to the Pearson chi-square statistic divided by the grand total of the table. The quality of the display is measured by the percentages of inertia accounted for by each principal axis, similar to PRINCIPAL COMPONENTS



Figure 1 Correspondence Analysis of Table 1: Symmetric Map in Two Dimensions

ANALYSIS. In Figure 1, the two axes explain a combined $70.5 + 14.7 = 85.2\%$ of the inertia, leaving 14.8% along further principal axes.

Although correspondence analysis applies primarily to CONTINGENCY TABLES, the method can also be used to visually interpret patterns in multivariate categorical data (multiple correspondence analysis), preferences, rankings, and ratings. This is achieved by appropriate transformation of the original data prior to applying the correspondence analysis procedure.

—Michael Greenacre

REFERENCES

- Benzécri, J.-P. (1973). *Analyse des Données. Tome 2: Analyse des Correspondances* [Data analysis: Vol. 2. Correspondence analysis]. Paris: Dunod.
- Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
- Greenacre, M. J. (1993). *Correspondence analysis in practice*. London: Academic Press.
- Greenacre, M. J., & Blasius, J. (1994). *Correspondence analysis in the social sciences*. London: Academic Press.

COUNTERBALANCING

Participants are exposed to all treatment conditions in a REPEATED-MEASURES DESIGN, which leads

to concerns about CARRYOVER EFFECTS and ORDER EFFECTS. Counterbalancing refers to exposing participants to different orders of the treatment conditions to ensure that such carryover effects and order effects fall equally on the experimental conditions.

Counterbalancing can be either complete or incomplete. With complete counterbalancing, every possible order of treatment is used equally often. Thus, with k treatments, there would be $k!$ different orders. As seen below, for three treatment conditions (A, B, and C), each treatment would occur equally often in each position. Moreover, each treatment would be preceded by every other treatment equally often (see Table 1).

Table 1

	1st Position	2nd Position	3rd Position
Order 1	A	B	C
Order 2	A	C	B
Order 3	B	A	C
Order 4	B	C	A
Order 5	C	A	B
Order 6	C	B	A

The advantage of counterbalancing is best understood by considering what might happen in the above experiment without counterbalancing. Suppose that all participants receive the treatments in the same order (A, B, C) and that mean performance under C is found to be significantly better than mean performance under A. Such an outcome cannot be interpreted

Table 2

Order	P1	P2	P3	P4	P5	Order	P1	P2	P3	P4	P5
1	A	B	E	C	D	6	D	C	E	B	A
2	B	C	A	D	E	7	E	D	A	C	B
3	C	D	B	E	A	8	A	E	B	D	C
4	D	E	C	A	B	9	B	A	C	E	D
5	E	A	D	B	C	10	C	B	D	A	E

unambiguously because of the CONFOUNDING of treatment and position. That is, the difference may be due to the treatment (A vs. C) *or* due to a practice effect (first position vs. third position), in which a participant's scores will improve over time regardless of treatment.

Counterbalancing the order of exposure to the treatment will eliminate the confounding of treatment and order. However, doing so will also inflate the error term in an analysis because any systematic order effects will increase variability of scores within each treatment condition. Order effects may be estimated and removed from the error term, as detailed elsewhere (cf. Keppel, 1991, pp. 361–373).

Complete counterbalancing is ideal because it ensures not only that each treatment occurs equally often in each position but also that each treatment occurs equally often in *all* of the preceding positions (P1, P2, etc.). However, with larger numbers of treatments, complete counterbalancing may be impractical. Under those circumstances, incomplete counterbalancing should be used.

One useful incomplete counterbalancing approach is referred to as a digram-balanced LATIN SQUARE. With this approach, when the number of treatments (k) is even, then only k orders are needed. When the number of treatments is odd, then $2 \times k$ orders are needed.

Each order is constructed in the following way. Lay out the treatment conditions in a row (ABCDE). The first order would be first, second, last, third, next to last, fourth, and so forth. To create the second order, first rearrange your row of conditions by moving the first treatment to the last position, with all other conditions remaining in the same order (BCDEA). Then select your order in the same way as before (first, second, last, etc.). Continue to rearrange and select in the same fashion until you have obtained k orders. When k is even, you would be finished when you had determined k orders.

The five orders on the left would result from this procedure for $k = 5$ (see Table 2). Every treatment occurs in every position equally often, which is important. However, because k is odd, treatments are not preceded by every other treatment. For example, as seen in Table 2 for Orders 1 through 5, B is never preceded by C or E. For that reason, when k is odd, it is essential that each order be mirrored, as seen in Orders 6 through 10. As you can see with this example, the digram-balanced approach ensures that every condition occurs in every position equally often and that every condition is preceded by every other condition equally often.

—Hugh J. Foley

REFERENCES

- Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Maxwell, S. E., & Delaney, H. D. (2000). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, NJ: Lawrence Erlbaum.

COUNTERFACTUAL

A counterfactual assertion is a conditional whose antecedent is false and whose consequent describes how the world *would have been* if the antecedent had obtained. The counterfactual takes the form of a subjunctive conditional: “If P had obtained, then Q would have obtained.” In understanding and assessing such a statement, we are asked to consider how the world would have been if the antecedent condition had obtained. For example, “If the wind had not reached 50 miles per hour, the bridge would not have collapsed” or “If the Security Council had acted, the war would have been averted.” We can ask two types of questions about counterfactual conditionals: What is the meaning of the statement, and how do we determine whether it is true or false? A counterfactual conditional cannot

be evaluated as a truth-functional conditional because a truth-functional conditional with false antecedent is ipso facto true (i.e., “if P then Q ” is equivalent to “either not P or Q ”). So is it necessary to provide a logical analysis of the truth conditions of counterfactuals if they are to be useful in rigorous thought.

There is a close relationship between counterfactual reasoning and causal reasoning. If we assert that “ P caused Q (in the circumstances C_i),” it is implied that we would assert the following: “If P had not occurred (in circumstances C_i), then Q would not have occurred.” So a causal judgment implies a set of counterfactual judgments. Symmetrically, a counterfactual judgment is commonly supported by reference to one or more causal processes that would have conveyed the world from the situation of the antecedent to the consequent. When we judge that the Tacoma Narrows Bridge would not have collapsed had the wind not reached 50 miles per hour, we rely on a causal theory of the structural dynamics of the bridge and the effects of the wind in reaching the consequent.

How do we assign a truth value to a counterfactual statement? The most systematic answer is to appeal to causal relations and causal laws. If we believe that we have a true causal analysis of the occurrence of Q , and if P is a necessary part of the set of sufficient conditions that bring Q to pass, then we can argue that, had P occurred, Q would have occurred. David Lewis (1973) analyzes the truth conditions and logic of counterfactuals in terms of possible worlds (possible world semantics). A counterfactual is interpreted as a statement about how things occur in other possible worlds governed by the same laws of nature. Roughly, “in every possible world that is relevantly similar to the existing world but in which the wind does not reach 50 miles per hour, the bridge does not collapse.” What constitutes “relevant similarity” between worlds is explained in terms of “being subject to the same laws of nature.” On this approach, we understand the counterfactual “If P had occurred, Q would have occurred” as a statement along these lines: “ P & {laws of nature} entail Q .” This construction introduces a notion of “physical necessity” to the rendering of counterfactuals: Given P , it is physically necessary that Q occurs.

—Daniel Little

REFERENCE

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

COVARIANCE

Covariance represents the fundamental measure of association between two random variables. Here we will discuss the concept of covariance in terms of the bivariate NORMAL DISTRIBUTION.

To begin, consider two random variables X and Y , representing, for example, a measure of reading proficiency and a measure of mathematics proficiency. Assume also that the two variables follow a bivariate normal distribution. Define $E(X)$ and $E(Y)$ to be the expectations of the random variables X and Y . Then, the covariance of X and Y , denoted as $\text{Cov}(X, Y)$ is

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)], \quad (1)$$

where $\mu_x = E(X)$ and $\mu_y = E(Y)$. A computation formula for the unbiased estimator of the population covariance can be written as

$$\frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}, \quad (2)$$

where N is the sample size, $\bar{X}$ is the sample mean of X , and $\bar{Y}$ is the sample mean of Y .

Returning to equation (1), the means of the bivariate normal distribution are contained in the mean vector μ , and the covariance of X and Y is contained in the *covariance matrix* Σ . The covariance matrix is a symmetric matrix, with the variances of X and Y on the diagonal and the covariance of X and Y on the off-diagonal.

The covariance represents the linear relationship between the random variables X and Y ; positive values of the covariance are likely to occur when $X - \mu_x$ and $Y - \mu_y$ are of the same sign, and negative values of the covariance are likely to occur when $X - \mu_x$ and $Y - \mu_y$ are of opposite sign. However, regardless of the sign of the covariance, its value is meaningless because it depends on the variances and hence on the metrics of the random variables themselves. To mitigate this problem, the CORRELATION coefficient standardizes the covariance to a mean of zero and a variance of 1 and hence puts the random variables on a common metric. This yields a measure of linear association that ranges from -1 to $+1$, with zero representing the absence of a linear relationship.

The covariance is a statistic and hence has a SAMPLING DISTRIBUTION. Imagine repeated draws

of the two random variables from a population characterized by a bivariate normal distribution. For each draw, we can calculate the means, variances, and the covariance of the two variables. The sampling distribution of the means is known to be normally distributed, and the sampling distributions of the variances are known to be CHI-SQUARE distributed. Concentrating on the covariance matrix, the sampling distribution of the covariance matrix is referred to as the *Wishart distribution*. The Wishart distribution is the multivariate analog of the chi-square distribution (see, e.g., Timm, 1975). It is interesting to note that the MAXIMUM LIKELIHOOD fitting function used for procedures such as FACTOR ANALYSIS and STRUCTURAL EQUATION MODELING can be derived from the Wishart distribution.

Having defined the covariance between two random variables and its sampling distribution, it is instructive to consider its use in applied settings. As noted earlier, the covariance provides a measure of linear association between two random variables. Also as noted, it is more common to transform the covariance into a correlation coefficient for ease of interpretation. Nevertheless, the covariance—particularly the covariance matrix—is a very important measure for other procedures. A large number of multivariate statistical procedures have been defined based on expressing the covariances in terms of more fundamental parameters. Such procedures as REGRESSION analysis, PATH ANALYSIS, factor analysis, and structural equation modeling can be considered as models for the covariances of the variables.

—David W. Kaplan

REFERENCE

Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology*. Monterey, CA: Brooks/Cole.

COVARIANCE STRUCTURE

The term *covariance structure* is used to designate a large class of statistical models that seek to represent the variances and covariances of a set of variables in terms of a smaller number of parameters. The premise here is that the sample variances and covariances are sufficient for the estimation of model parameters. The term *mean structure* is used to denote a class of models that represent the means of the sample data in terms of a smaller number of parameters. The general problem

of covariance structure modeling has been taken up in Browne (1982).

Models that are subsumed under the term *covariance structure* include but are not limited to regression analysis, path analysis, factor analysis, and structural equation modeling. The general form of the model can be written as

$$\Sigma = \Sigma(\Theta), \quad (1)$$

where Σ is a $p \times p$ population covariance matrix with p being the number of variables, and Θ is a q -dimensional vector of model parameters with $q \leq p$. The function in equation (1) states that the covariance matrix can be expressed as a matrix-valued function of a parameter vector.

As stated above, the covariance structure model in equation (1) subsumes a number of commonly used statistical procedures. What is required is that the parameters of the model can be expressed in terms of elements of the covariance matrix elements. Consider the simple example of a two-variable linear regression model. Denote the model as

$$y = x\beta + u, \quad (2)$$

where y and x are the outcome and regressor variables, respectively; β is the regression coefficient; and u is the regression disturbance term. Equation (2) can be seen as a covariance structure when the regression coefficient β can be estimated as

$$\hat{\beta} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}, \quad (3)$$

where $\text{Cov}(x, y)$ is the covariance of x and y , and $\text{Var}(x)$ is the variance of x . Similarly, the variance of the disturbance term u (denoted as σ_u^2) can be solved in terms of elements of Σ . Thus, we see that regression analysis is a special case of covariance structure modeling.

To take another example of the generality of the covariance structure representation, consider the factor analysis model, written in linear form as

$$\mathbf{y} = \Lambda\boldsymbol{\eta} + \boldsymbol{\zeta}, \quad (4)$$

where $\mathbf{y}$ is a $p \times 1$ vector of variables, Λ is a $p \times k$ matrix of factor loadings, $\boldsymbol{\eta}$ is a $k \times 1$ vector of common factors, and $\boldsymbol{\zeta}$ is a $p \times 1$ vector of uniquenesses (composed of measurement error and specific error). Assuming that the mean of the factors and the uniquenesses are zero and that the covariance of the factors and the

uniquenesses are zero, the covariance matrix for $\mathbf{y}$ can be expressed as

$$\Sigma_{\mathbf{y}} = \Lambda \Phi \Lambda' + \Psi, \quad (5)$$

where Φ is the matrix of covariances among the factors, and Ψ is a diagonal matrix of unique variances. It can be seen from equation (5) that the covariance matrix is expressed in terms of the fundamental parameters of the factor analysis model and hence is, by definition, a covariance structure.

Finally, it may be of interest to see how structural equation modeling fits into the general covariance structure framework. The system of structural equations can be written as

$$\mathbf{y} = \alpha + \mathbf{B}\mathbf{y} + \Gamma\mathbf{x} + \zeta, \quad (6)$$

where $\mathbf{y}$ is a $p \times 1$ vector of observed endogenous variables, $\mathbf{x}$ is a $q \times 1$ vector of observed exogenous variables, α is a $p \times 1$ vector of structural intercepts, $\mathbf{B}$ is a $p \times p$ coefficient matrix that relates endogenous variables to each other, Γ is a $p \times q$ coefficient matrix that relates endogenous variables to exogenous variables, and ζ is a $p \times 1$ vector of disturbance terms where $\text{Var}(\zeta) = \Psi$ is the $p \times p$ covariance matrix of the disturbance terms, where $\text{Var}(\cdot)$ is the variance operator. Finally, let $\text{Var}(\mathbf{x}) = \Phi$ be the $q \times q$ covariance matrix for the exogenous variables.

The system described in equation (6) is referred to as the *structural form* of the model. It is convenient to rewrite the structural model so that the endogenous variables are on one side of the equation and the exogenous variables on the other side. Thus, equation (6) can be rewritten as

$$(\mathbf{I} - \mathbf{B})\mathbf{y} = \alpha + \Gamma\mathbf{x} + \zeta. \quad (7)$$

Assuming that $(\mathbf{I} - \mathbf{B})$ is nonsingular so that its inverse exists, equation (7) can be written as

$$\begin{aligned} \mathbf{y} &= (\mathbf{I} - \mathbf{B})^{-1}\alpha + (\mathbf{I} - \mathbf{B})^{-1}\Gamma\mathbf{x} + (\mathbf{I} - \mathbf{B})^{-1}\zeta \\ &= \Pi_0 + \Pi_1\mathbf{x} + \zeta^*, \end{aligned} \quad (8)$$

where Π_0 is the vector of reduced-form intercepts, Π_1 is the vector of reduced-form slopes, and ζ^* is the vector of reduced-form disturbances with $\text{Var}(\zeta^*) = \Psi^*$. This specification is referred to as the *reduced form* of the model. Note that equation (8) is a straightforward multivariate regression of $\mathbf{y}$ on $\mathbf{x}$, and so this model also falls into the general form of a covariance structure.

Next, let Θ be a vector that contains the structural parameters of the model—so in this case, $\Theta = (\mathbf{B}, \Gamma, \Psi, \Phi)$. Furthermore, let $E(\mathbf{x}) = \mu_{\mathbf{x}}$ be the vector of means for $\mathbf{x}$, $\text{Var}(\mathbf{x}) = E(\mathbf{x}'\mathbf{x}) = \Phi$, and $E(\zeta) = \mathbf{0}$, where $E(\cdot)$ is the expectation operator. Then,

$$\begin{aligned} E(\mathbf{y}) &= (\mathbf{I} - \mathbf{B})^{-1}\alpha + (\mathbf{I} - \mathbf{B})^{-1}E(\mathbf{x}) \\ &= (\mathbf{I} - \mathbf{B})^{-1}\alpha + (\mathbf{I} - \mathbf{B})^{-1}\mu_{\mathbf{x}} \end{aligned} \quad (9)$$

and

$$\begin{aligned} E(\mathbf{y}, \mathbf{x}) = \Sigma_{\mathbf{yx}} &= \begin{bmatrix} E(\mathbf{y}\mathbf{y}') & E(\mathbf{y}\mathbf{x}') \\ E(\mathbf{x}'\mathbf{y}) & E(\mathbf{x}'\mathbf{x}) \end{bmatrix}, \\ &= \begin{bmatrix} (\mathbf{I} - \mathbf{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)(\mathbf{I} - \mathbf{B})^{-1} & (\mathbf{I} - \mathbf{B})^{-1}\Gamma\Phi \\ \Phi\Gamma'(\mathbf{I} - \mathbf{B})^{-1} & \Phi \end{bmatrix}. \end{aligned} \quad (10)$$

Thus, equations (9) and (10) show that structural equation modeling represents a structure on the mean vector and covariance matrix. The structure is in terms of the parameters of the model. Covariance structure modeling concerns the elements of equation (10).

It is clear from the above discussion that the covariance structure representation of statistical modeling has wide generality. The covariance structure representation also provides a convenient way to consider problems in model identification. Although model identification is taken up elsewhere in this encyclopedia, a necessary condition for the identification of model parameters is that the number of model parameters be less than or equal to the number of elements in the covariance matrix. Specifically, arrange the unknown parameters of the model in the vector Θ . Consider next a population covariance matrix Σ whose elements are population variances and covariances. It is assumed that an underlying, substantive model can be specified to explain the variances and covariances in Σ . We know that the variances and covariances in Σ can be estimated by their sample counterparts in the sample covariance matrix $\mathbf{S}$ using straightforward formulas for the calculation of sample variances and covariances. Thus, the parameters in Σ are identified.

Having established that the elements in Σ are identified from the sample counterparts, what we need to determine is whether the model parameters are identified. We say that the elements in Ω are *identified* if they can be expressed uniquely in terms of the elements

of the covariance matrix Σ . If all elements in Θ are identified, we say that the model is identified.

—David W. Kaplan

REFERENCE

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, UK: Cambridge University Press.

COVARIATE

In statistical analysis, the term *covariate* generally refers to the INDEPENDENT VARIABLE. For example, in the analysis of EXPERIMENTS, a covariate is an independent variable not manipulated by the experimenter but still affecting the response variable. In SURVIVAL or EVENT HISTORY ANALYSIS, a covariate is simply an independent variable, continuous or not, that may have an effect on the HAZARD RATE, such as a time-varying (or time-dependent) covariate. However, in LOG-LINEAR MODELING, continuous variables are treated as “cell covariates” by imposing a metric on categorical variables, instead of allowing use of truly continuous variables.

—Tim Futing Liao

COVER STORY

In some laboratory research, the particular hypotheses under investigation are relatively obvious. This transparency of research purpose may prove problematic if participants are inclined to assist the researcher by helping to confirm the apparent hypotheses. For example, suppose participants were asked to complete a questionnaire that asked a series of questions regarding their feelings toward citizens of Uzbekistan. Following this task, participants are shown a video of heroic Uzbekistanis defending their homeland against terrorists. Then, the attitude questionnaire is readministered. In this case, the goal of the research appears obvious. Research on DEMAND CHARACTERISTICS suggests that the apparently obvious aim of the study may produce artifactual results. Some participants will have determined that the purpose of the research is to induce them to indicate positive

evaluations of Uzbekistanis and will cooperate, even if the video (the INDEPENDENT VARIABLE) had little real effect on their feelings.

To offset the possibility of distorted results brought on by overly cooperative participants, psychologists have developed more or less elaborate cover stories designed to hide the true purpose of their research. In the example, the pretest might be omitted, and the participants are informed that their task is to judge the technical quality of the video they are to see. Then, after completing this task, they might be administered a posttest that taps their feelings toward the Uzbekistanis. As an addition to the cover story, this task might be introduced in a way that also hides the true nature of the research. These participants' responses could then be compared in a *posttest-only/comparison group design* with those of participants who saw no video or who saw a different video (Crano & Brewer, 2002).

Cover stories are used to disguise the true purpose of research and hence are deceptive. Thus, their ethical use must be considered carefully. We justify cover stories because they are used not principally to deceive but to ensure the validity of research findings at little cost to participants. Typically, they are used in LABORATORY EXPERIMENTS, where the danger of misleading findings due to overly cooperative participants is most severe.

Crano and Brewer (2002) distinguished among three general classes of participants in research: voluntary, involuntary, and nonvoluntary. Nonvoluntary participants, by definition, are unaware they are the subject of scientific investigation, so there is no need for a cover story when using such participants. Voluntary and involuntary participants, however, are well aware of their status. Voluntary participants are content to serve in research, and their positive attitudes may prove dangerous to validity if they motivate attempts to help the investigator confirm research hypotheses. Involuntary respondents also are aware of their role but are unhappy about it, often feeling coerced into participating. These participants' negative attitudes may evolve into attempts to ruin the research; in such cases, a cover story may prove useful because it helps obscure the true intent of the research.

—William D. Crano

REFERENCE

Crano, W. D., & Brewer, M. B. (2002). *Principles and methods of social research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

COVERT RESEARCH

Covert research is that which proceeds without giving subjects or their legitimate representatives the basic information that they are being researched. All the professional associations relating to social research reflect in their ethical guidelines the principle of INFORMED CONSENT, and they advise uniformly against covert methods.

ETHICAL ISSUES

There are compelling reasons why social researchers should want to avoid covert practice.

First and foremost, there is an issue of autonomy. Subjects are accorded a right to be informed, to control their own observed behavior, and to make their own decisions on the risks or benefits of participation. Moreover, the data they supply have a value, and they may be entitled to put a price on what researchers take from them; researchers who do not inform subjects that they are yielding these data may be thought to be plundering or stealing them.

Furthermore, the use of methods that are not open will get researchers a bad name. At one time, Henle and Hubble (1938) stowed themselves under the beds in student dorms to record conversations; later, Laud Humphreys (1975) posed as a lookout or “watchqueen” to observe homosexual encounters in public toilets. Nowadays, those who research human behavior resist the reputation of eavesdropper or voyeur. The professional associations in the social and behavioral sciences claim the moral high ground and prefer to leave covert research to investigative journalists. They recognize also that behavior may be private, even though the territory in which it is observable may be public.

JUSTIFICATIONS

The consensus of professional associations and codes notwithstanding, there persists a debate concerning the acceptability of covert methods (Bulmer, 1982; Homan, 1992). It is not realistic to insist that all subjects should know that they are being observed and be given the option to withdraw. Some of the most perceptive work in the social sciences has taken the form of retrospective analysis of everyday experience. Erving Goffman interpreted the putting down

of towels and newspapers on beaches and trains as territorial behavior: No one would ask that consent be sought or that a notice should be posted saying, “Caution: this beach is being observed by a sociologist.” Once possessed of what Wright Mills (1959) called the “sociological imagination,” students of human behavior become habitual researchers: The act of research is not the collection of data but the subsequent construction of them, against which there is no law.

The right to know that research is taking place and therefore to decline or avoid it is itself problematic. It is, for example, accorded more to the living than to the dead whose papers turn up in county record offices. The right to refuse research may run contrary to the public interest. For example, it is arguable that social researchers should complement the investigations of journalists and police to expose places where children or other vulnerable groups are being abused or where racist attacks are being sponsored. Indeed, the intellectual freedom and integrity of social researchers are compromised if they are prepared to take “no” for an answer. The pity is that subjects in positions of power are more prone to insulate themselves from open investigators by withholding consent than are the relatively defenseless and disempowered; in consequence, the ethic of openness discriminates against the weak.

Covert methods are, in some circumstances, more likely to be nonreactive. This is true in two senses. First, because subjects will not know that they are on the record, they will not have reason to adjust behavior or data for the sake of a favorable report. Second, the research act will not inhibit or devalue their discourse: Participants in an intimate prayer meeting need to be unselfconscious, and it would therefore be disturbing for them to know that one of their company was only there to observe. One of the ways of overcoming the issue of REACTIVITY with informed subjects is to stay in the field for a sufficient time until one becomes accepted as part of the physical setting; another common strategy is to use a gatekeeper or informed collaborator who can speak for the group while not being a member of it.

COMMENTARY

It is not possible to draw a clear line between covert and open methods. Even those who use the

most explicit of methods, such as interviews and questionnaires, deploy skills to redefine the research act and to diminish their subjects' awareness of it. Again, they will ask informants questions about other persons whose consent has not been sought, thus using interviewees as surrogate informants while not physically engaging those who are the effective subjects in the research. Although being ethical in the sense of being in accordance with the professional codes, open methods are not necessarily more moral.

What weighs even more heavily than the argument against covert methods is the professional consensus. The principle of informed consent is that on which the self-respect and reputation of social research rest, and practitioners defy it at their peril.

More than 10 years ago, the standard objections to covert practice were that it constituted an infringement of personal liberty and a betrayal of trust in a world, at a time when these things were highly valued. Recent years have seen the advent of security cameras in public places, computer records and databases that can be hacked into by those with appropriate skills, and cases of dishonesty and sleaze in high places. The milieu of trust and liberty has changed. Social researchers endeavor to convey findings that politicians then spin. It may well be that the credibility of researchers will stand or fall not by the openness of the methods they deploy but by the extent to which they can bypass control and manipulation and offer findings that users can trust.

—Roger Homan

REFERENCES

- Bulmer, M. (Ed.). (1982). *Social research ethics*. London: Macmillan.
- Henle, M., & Hubble, M. B. (1938). Egocentricity in adult conversation. *Journal of Social Psychology*, 9, 227–234.
- Homan, R. (1992). *The ethics of social research*. London: Longman.
- Humphreys, L. (1975). *Tearoom trade: Impersonal sex in public places*. Chicago: Aldine.
- Mills, C. W. (1959). *The sociological imagination*. New York: Oxford University Press.

CREATIVE ANALYTICAL PRACTICE (CAP) ETHNOGRAPHY

From the 1980s onward, influenced by POST-MODERNISM, feminism, and POSTSTRUCTURALISM, contemporary ETHNOGRAPHERS have challenged standard social science writing as ethically, philosophically, and scientifically problematic. Recognizing that the boundaries between “fact” and “fiction,” “subjective and objective,” and “author and subject” are blurred, these contemporary ethnographers are writing social science in new ways. At first, this work was referred to as *experimental ethnography* or *alternative ethnography*. More recently, to underscore the processes by which the writing is accomplished and to establish the legitimacy of these ethnographies in their own right, they are increasingly known as *creative analytical practice* or *CAP ethnography*.

CAP ethnography is both “scientific”—in the sense of being true to a world known through empirical work and study—and “literary”—in the sense of expressing what one has learned through evocative writing techniques and forms. Although CAP ethnographers belong to a number of different disciplines, they tend to share some research/writing practices in common. These include using theories that challenge the grounds of disciplinary authority; writing on topics of social and personal importance; privileging nonhierarchical, collaborative, and/or multivocal writing; favoring self-reflexivity; positioning oneself in multiple, often competing discourses; and the signature practice, writing evocatively, for different audiences in a variety of formats.

The varieties of CAP ethnography are many (cf. Richardson 1997). They include AUTOETHNOGRAPHY (cf. Ellis, 2003), FICTION, poetic representation, performance texts, ethnographic drama, story writing, reader's theater, aphorisms, conversations, epistles, mixed media, and layered accounts, as well as hypertexts, museum exhibits, photography, paintings, dance, and other forms of artistic representation. For example, Laurel Richardson created a narrative poem from an open-ended interview with “Louisa May,” an unwed mother (see Richardson, 1997). Carolyn Ellis (2003) invented a fictional graduate seminar in which characters, based on composites of actual students, engage in dialogue about autoethnography.

EVALUATION OF CAP ETHNOGRAPHIES

Most CAP ethnographers subscribe to six criteria:

1. *Substantive*: Does the piece make a substantive contribution to our understanding of social life?
2. *Aesthetic*: Does the piece invite interpretive responses? Is it artistically satisfying and complex?
3. *REFLEXIVITY*: Is the author reflexive about the self as a producer of the text?
4. *ETHICS*: Does the author hold himself or herself accountable to the standards of knowing and telling of the people studied?
5. *Impact*: Does the text affect the reader? Emotionally? Intellectually? Move others to social actions? Did the text change the writer?
6. *Expresses a reality*: Does the text provide a credible account of a cultural, social, individual, or communal sense of the “real”?

CAP ethnographers recognize that the product cannot be separated from the producer—the problems of subjectivity, authority, authorship, and reflexivity are intricately interwoven into the representational form. Science is one lens, creative arts another. CAP ethnographers try to see through both lenses.

—Laurel Richardson

REFERENCE

- Denzin, N., & Lincoln, Y. (Eds.). (2000). *Handbook of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.
- Ellis, C. (2003). *The ethnographic “I”: A methodological novel on doing autoethnography*. Walnut Creek, CA: AltaMira.
- Richardson, L. (1997). *Fields of play: Constructing an academic life*. New Brunswick, NJ: Rutgers University Press.

CRESSIE-READ STATISTIC

The Cressie-Read statistic belongs to the family of power divergence statistics for assessing model fitting and is an alternative to the LIKELIHOOD RATIO STATISTIC, although it does not have a *p* value associated with the statistic.

—Tim Futing Liao

REFERENCE

- Cressie, N., & Read, T. (1984). Multinomial goodness of tests. *Journal of Royal Statistical Society Series B*, 46, 440–464.

CRITERION RELATED VALIDITY

In educational assessment and social science measurement, criterion-related validity is defined within the context of test or measurement validation. Test and measurement validation is the process whereby data are collected to establish the credibility of inferences to be made from test scores or measurements—both inferences about use of scores and measures (as in decision making) and inferences about interpretation (meaning) of scores and measures. Prior to the 1985 *Standards for Educational and Psychological Testing* (American Educational Research Association, 1985), validity was conceptualized, in the behaviorist tradition, as a taxonomy of three major and largely distinct categories (i.e., content validity, criterion-related validity, and CONSTRUCT VALIDITY). These three validation approaches provide accumulative evidence for a test’s or a measurement instrument’s validity. Criterion-related validity is evidence of a test’s or an instrument’s association with a criterion. This validation procedure is commonly used for aptitude tests and for other situations in which test scores or measurements are used for prediction. Criterion-related validity is further divided into two subtypes: (a) predictive evidence of criterion-related validity (or predictive validity) and (b) concurrent evidence of criterion-related validity (or concurrent validity). The predictive evidence of criterion-related validity indicates the effectiveness of a test or an instrument in predicting a test taker’s future performance in specified activities. For example, to claim that a harp aptitude test has predictive validity, test developers must demonstrate that high scores on this test are associated with superior performance on the harp, and low scores are associated with poor performance. The concurrent evidence of criterion-related validity indicates the test’s degree of association with a current criterion. Concurrent validity differs from predictive validity only in that the criterion is concurrent. For example, a test designed to diagnose psychological depression may be validated by administering it to persons whose level of depressive symptomatology is already known. Often, concurrent validity is used as a substitute for predictive validity because there is no

delay in the availability of criterion measures. So, for instance, if a scale is constructed to assess students' science aptitude, the concurrent validity may be established by correlating test scores with students' grades in a current science course, instead of with future indicators of science aptitude, such as accomplishments in science, writings on science topics, and so forth. The difference between these two validity evidences lies in the objective of a test. If a test or an instrument is to be used for predicting a criterion trait or behavior, the predictive validity is called for. If a test or an instrument is to be used to identify a criterion trait or behavior, the concurrent validity is used to make the validity claim.

The current thinking about the conceptualization of validity, as reflected in the 1985 and 1999 *Standards for Educational and Psychological Testing*, stresses that it is not the tests that are valid (or not); rather, inferences made about the meaning and the use of test scores are valid (or not). "Validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (American Educational Research Association, 1999, p. 9). Furthermore, validity is said to be a "unified though faceted concept" (Messick, 1989) for which different types of evidence (not different types of validity) should be gathered to support the inferences made from test scores (or measurements) and proper use of the test (or the instrument). The instances of criterion-related validity evidence discussed above are examples of such types of evidence.

—Chao-Ying Joanne Peng and Daniel J. Mueller

See also CONSTRUCT VALIDITY, VALIDITY

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31–41.
- Messick, S. (1989, March). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, pp. 5–11.
- Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (6th ed., pp. 1487–1495). New York: Macmillan.

CRITICAL DISCOURSE ANALYSIS

Critical discourse analysis (CDA) sees language as one element of social events and social practices that is dialectically related to other elements (including social institutions and aspects of the material world). Its objective is to show relationships between language and other elements of social events and practices. It is an approach to language analysis that originates within linguistics but is more socially oriented and interdisciplinary than most linguistics.

Discourse here means language (as well as "body language," visual images, etc.) seen as a moment of the social. The approach is "critical" in the sense that (a) it seeks to illuminate nonobvious connections between language and other elements of social life; (b) it focuses on how language figures in the constitution and reproduction of social relations of power, domination, and exploitation; and (c) it takes a specifically language focus on social emancipation and the enhancement of social justice. In these respects, it can be seen as a branch of critical social science.

CDA is based on the assumption that the language elements of social events (talk, texts) can contribute to change in other social elements—that discourse is socially constructive. A focus of analysis has been on the effects of discourse in constituting, reproducing, and changing ideologies. Both are consistent with a dialectical view of discourse as an element of the social that is different from others while not being discrete—different elements "internalize" each other. One aspect of recent changes in social life is arguably that discourse has become more salient in certain respects in relation to other elements; for instance, the concept of a "knowledge" or "information" society seems to imply that social (e.g., organizational) change is "led" by discourses that may be enacted in new ways of acting, inculcated in new identities, materialized in new plants, and so forth. However, in contrast with a certain discourse idealism, we should take a moderate view of social CONSTRUCTIVISM, recognizing that

social constructs acquire intransitive properties that may make them resistant to the internalization of new discourses.

DEVELOPMENT OF THE FIELD

Critical perspectives on the relationship of language to other elements of social life can be traced back to Aristotelian RHETORIC and found across a range of academic disciplines. Although CDA is not original in addressing such issues, it has developed for the first time a relatively systematic body of theory and research. Its main sources include broadly “functionalist” (as opposed to “formalist”) traditions within linguistics, including the systemic functional linguistics theory of Michael Halliday; theorizations of “hegemony” and “ideology” within Western Marxism, notably by Gramsci and Althusser; the “critical theory” of the Frankfurt School, including more recently Habermas; the concept of “discourse” in Foucault; and the dialogical language theory of Bakhtin, including its emphasis on “heteroglossia” and “intertextuality” and its theorization of “genre.”

The term *critical discourse analysis* was first used around 1985, but CDA can be seen as including a variety of approaches, beginning with *critical linguistics*, an application of Halliday’s linguistics to analyzing texts from the perspective of ideology and power. Some work in CDA (especially that of Teun van Dijk) incorporates a cognitive psychology, some includes an emphasis on historical documentation (especially the “discourse-historical” approach of Ruth Wodak), and some (especially Gunther Kress and Theo van Leeuwen) has focused on the “multi-modal” character of contemporary texts, particularly the mixture of language and visual image. CDA has progressively become more interdisciplinary, engaging more specifically with social theory and research, and (especially in the case of Fairclough) it is committed to enhancing the capacity of research on the social transformations of the contemporary world (globalization, neoliberalism, new capitalism) to address how language figures in processes of social transformation.

APPLICATION

CDA has recently been drawn on in research within a considerable variety of disciplines and areas of study in the social sciences (e.g., geography and urban

studies, health sciences, media studies, educational research, sociology). A major division within this extensive uptake is between studies that incorporate detailed linguistic analysis of texts and studies that work with a generally Foucaultian view of discourse and draw on theoretical perspectives in CDA without such close textual analysis. For some CDA researchers, establishing the value of close textual analysis in social research is a central concern. At the same time, there is a concern to contribute a language focus to social and political debates in the public sphere (Fairclough, 2000).

EXAMPLE

A theme that has attracted considerable international attention and now has its own virtual research network is Language in New Capitalism (www.cddc.vt.edu/host/lnc/), which explores how language figures in the current restructuring and “global” rescaling of capitalism. An example of what CDA can contribute to social research on these transformations is analysis of narratives of global economic change (Fairclough, 2000). Such narratives are pervasive in economic, political, and other types of texts. Economic change is recurrently represented within them as a process that does not involve human agents—so, for instance, new markets “open up” rather than people opening them up, and technologies “migrate” from place to place rather than being transported by companies—which is universal in time and place (it happens in a universal present and everywhere equally). Such representations of economic change are widely worked together with policy injunctions—what “is” unarguably and inevitably the case dictates what we “must” do in policy terms. Analysis of these narratives can put some flesh on the rather skeletal argument advocated by Bourdieu that neoliberal discourse is a powerful part of the armory for bringing into existence a “globalization” (and transformation of capitalism) that is misleadingly represented as already fully achieved.

—Norman Fairclough

REFERENCES

- Chouliaraki, L., & Fairclough, N. (1999). *Discourse in late modernity*. Edinburgh, UK: Edinburgh University Press.
- Fairclough, N. (2000). *New labour, new language?* London: Routledge Kegan Paul.

- Fairclough, N., & Wodak, R. (1997). Critical discourse analysis. In T. van Dijk (Ed.), *Discourse as social interaction*. Thousand Oaks, CA: Sage.
- Wodak, R., & Meyer, M. (2001). *Methods in critical discourse analysis*. Thousand Oaks, CA: Sage.

CRITICAL ETHNOGRAPHY

In recent years, critical ethnography has become an increasingly popular perspective.

However, how scholars from different disciplines have tailored it by using different theoretical or paradigmatic approaches has created a diverse range of applications that result in labeling any form of cultural criticism as *critical ethnography*. Therefore, it may be less important to define critical ethnography than to identify a few fundamental constituent elements.

Reducing critical ethnography simply to social criticism distorts and oversimplifies the critical ethnographic project. At its simplest, critical ethnography is a way of applying a subversive worldview to more conventional narratives of cultural inquiry. It does not necessarily stand in opposition to conventional ETHNOGRAPHY or even to conventional social science. Rather, it offers a more reflective style of thinking about the relationship between knowledge, society, and freedom from unnecessary forms of social domination.

What distinguishes critical ethnography from the other kind is not so much an act of criticism but an act of critique. *Criticism*, a complaint we make when our eggs are undercooked, generally connotes dissatisfaction with a given state of affairs but does not necessarily carry with it an obligation to dig beneath surface appearances to identify fundamental social processes and challenge them.

Critique, by contrast, assesses “how things are” with an added premise that “things could be better” if we examine the underlying sources of conditions, including the values on which our own complaints are based.

Critical ethnography is not a theory because it does not, in itself, generate a set of testable propositions or lawlike statements. Rather, it is a perspective that provides fundamental images, metaphors, and understandings about our social world. It also provides value premises around which to ask questions and suggest ways to change society. It is a methodology that “strives to unmask hegemony and address oppressive forces” (Crotty, 1998, p. 12). Critical researchers begin

from the premise that all cultural life is in constant tension between control and resistance. This tension is reflected in behavior, interaction rituals, normative systems, and social structure, all of which are visible in the rules, communication systems, and artifacts that comprise a given culture. Critical ethnography takes seemingly mundane events, even repulsive ones, and reproduces them in a way that exposes broader social processes of control, power imbalance, and the symbolic mechanisms that impose one set of preferred meanings or behaviors over others. Although not all critical ethnographers are equally comfortable with issues such as GENERALIZABILITY, RELIABILITY, VALIDITY, and VERIFICATION, critical ethnography is quite compatible with the conventions of “empirical science” by which the credibility of empirical claims can be evaluated (Maines, 2001).

Two examples illustrate the diversity of critical ethnography. First, Bosworth’s (1999) critical ethnography of culture in three English women’s prisons identifies a paradox: The conventional idealized images of heterosexual femininity that function as part of the punishment process also become a means for women to resist control and survive their prison experience. She argues that, although an idealized notion of femininity underlies much of the daily routine of women’s prisons, it possesses the contradictory and ironic outcome of producing the possibility for resistance (Bosworth, 1999, p. 107). This type of critique helps soften the jagged edges of the gender gap in criminal justice scholarship by reversing the longstanding tradition of looking at women’s prison culture through the prism of male prisoners and male scholars.

Ferrell and Hamm (1998) provide a collection of insightful examples that further demonstrate the utility of critical ethnography. By looking at marginalized populations such as pimps, marijuana users, phone sex workers, and domestic terrorists, the contributors challenge the reader to recast conventional, comfortable assumptions and images about crime, deviance, and social control. This, in turn, challenges us to examine more fully the asymmetrical power relations that create and sustain sociocultural power imbalances.

There are, of course, problems with critical ethnography (Thomas, 1992). But science—and critique is part of the scientific project—is self-correcting. One task of those engaged in ethnographic critique lies in balancing the occasionally conflicting tasks of a priori conceptual analysis, interpretive narration, empirical rigor, and theory building, taking care not to reject such

tasks as POSITIVIST lest we substitute one intellectual dogma with another. This means that we must continually reassess our own project with the same rigor that we assess our foci of analysis, always bearing in mind that things are never what they seem and that social justice is not simply a goal but a vocation to which all scholars ought strive with verifiable critique.

—Jim Thomas

REFERENCES

- Bosworth, M. (1999). *Engendering resistance: Agency and power in women's prisons*. Aldershot, UK: Ashgate/Dartmouth.
- Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. Thousand Oaks, CA: Sage.
- Ferrell, J., & Hamm, M. S. (1998). True confessions: Crime, deviance, and field research. In J. Ferrell & M. Hamm (Eds.), *Ethnography at the edge: Crime, deviance and field research* (pp. 2–19). Boston: Northeastern University Press.
- Kirk, J., & Miller, M. L. (1986). *Reliability and validity in qualitative research*. Beverly Hills, CA: Sage.
- Maines, D. R. (2001). *The faultline of consciousness: A view of interactionism in sociology*. New York: Aldine De Gruyter.
- Thomas, J. (1992). *Doing critical ethnography*. Newbury Park, CA: Sage.

CRITICAL HERMENEUTICS

Critical hermeneutics grows out of the hermeneutic phenomenology of Paul Ricoeur (1981; Thompson, 1981) and the depth hermeneutics of John Thompson (1990). Following Ricoeur, it decomposes interpretation into a dialectical process involving three interrelated, interdependent, mutually modifying phases or “moments”: a moment of social-historical analysis, a moment of formal analysis, and a moment of interpretation-reinterpretation in which the first two moments of analysis are brought together (see Phillips & Brown, 1993, for an empirical example). In interpreting a text, one can and should analyze the social-historical context out of which it has arisen. But one should also analyze the text formally, as a system of signs abstracted from experience using one of the formal approaches to textual analysis such as structural SEMIOTICS or DISCOURSE ANALYSIS. From a critical hermeneutic perspective, neither the analysis of the social-historical context nor the text itself is superior. Each is an indispensable moment in the interpretive process, and methods of interpretation that fail to

explicitly include the social-historical context *and* the formal analysis of the text are incomplete. Having completed the social-historical and formal analysis, a researcher must creatively combine these two moments to produce an interpretation of the text and its role in the social system of which it is a part. It is in this moment of interpretation that the researcher develops insights into the role of the text in the ongoing construction of social reality. This decomposition of the process into moments provides a structured framework for the researcher to follow and allows readers to more clearly understand the researcher's methodology and to judge more clearly the value of the final conclusions.

But in addition to being hermeneutic, the methodology is also critical. In addition to being concerned with meaning, critical hermeneutics is concerned with the role of communication in the ongoing maintenance of the asymmetrical power relations that characterize the social world. The examination of the social-historical context of the production, transmission, and reception of a text provides an opportunity for researchers to include a concern for the role of communication in the development of relations of power that lead to enduring asymmetries of opportunities and life chances. It allows for the introduction of an emancipatory interest and focuses attention on the taken-for-granted conventions and rules that underlie a particular social system. Those applying a critical hermeneutic methodology ask how the communication process in a certain social setting functions, for whose benefit it functions, and who is allowed to participate.

By combining an interest in interpretation of texts with the social-historical context of the text, critical hermeneutics makes three important contributions. First, it provides a structured method for examining the role of symbolic phenomena in social life. Second, it provides a structured method for examining the sources of texts and the interests of their producers. Third, it provides a way to integrate formal methods of textual analysis with an interest in the social context in which the texts occur. Critical hermeneutics therefore adds an important dimension to more traditional interpretive methods in social science.

—Nelson Phillips

REFERENCES

- Phillips, N., & Brown, J. (1993). Analyzing communication in and around organizations: A critical hermeneutic approach. *Academy of Management Journal*, 36(6), 1547–1576.

- Ricoeur, P. (1981). *Hermeneutics and the human sciences: Essays on language, action and interpretation*. New York: Cambridge University Press.
- Thompson, J. (1981). *Critical hermeneutics: A study in the thought of Paul Ricoeur and Jurgen Habermas*. Cambridge, UK: Cambridge University Press.
- Thompson, J. (1990). *Ideology and modern culture*. Stanford, CA: Stanford University Press.

CRITICAL INCIDENT TECHNIQUE

The critical incident technique (CIT) was first devised and used in a scientific study almost half a century ago (Flanagan, 1954). The method assumed a POSITIVIST approach, which was at the time the dominant paradigm in the social as well as the natural sciences. It was used principally in occupational settings, and its advocates established the VALIDITY and RELIABILITY of the technique (Andersson & Nilsson, 1964; Ronan and Latham, 1974). The approach of Flanagan and the studies he reviewed assumed tangible reality. In contrast, CIT, as developed by Chell (1998), assumes PHENOMENOLOGY. It is intended through the method of, technically, an UNSTRUCTURED INTERVIEW to capture the thought processes, the frames of reference, and the feelings about an incident that have meaning for the RESPONDENT. In the interview, respondents are required to give an ACCOUNT of what the incident means for them in relation to their life situation, as well as their present circumstances, attitudes, and orientation. The approach may be used in an in-depth CASE STUDY, as well as a multisite investigation.

CIT is a QUALITATIVE INTERVIEW procedure, which facilitates the investigation of significant occurrences (events, incidents, processes, or issues) identified by the respondent, the way they are managed, and the outcomes in terms of perceived effects. The objective is to gain an understanding of the incident from the perspective of the individual, taking into account cognitive, affective, and behavioral elements (Chell, 1998, p. 56).

EXAMPLE AND METHOD

Arguably, all research commences with “preliminary design work.” This assumes that there is a point to the research; the research questions and ideas are coherent and focused on the aims of the study. Moreover, in empirical work, the researcher

needs to consider how to locate the respondent (or subject) and why this particular respondent was chosen. Researchers may use a variety of sources from their own local knowledge, word of mouth, and databases and lists owned by other bodies. This also means that there is likely to be a criterion or criteria for inclusion of the subject that need to be made explicit. An example in which the use of the CIT was helpful in exposing a depth of rich information was a study of personal and business issues on microbusinesses in business services. It showed the importance of personal and family issues relative to other business and economic factors (see Chell, 1998; Chell & Baines, 1998).

Hand in hand with locating subjects is the ability to gain ACCESS to them. This may require skill, patience, and perseverance. It is wise not to take access for granted. Devices such as an initial introductory telephone call or letter of introduction may be used. The latter approach enables the researcher to outline the aims and objectives of the project and offer plausible reasons for seeking to intrude on the subject’s time and PRIVACY. Hence, sensitive handling is required from the outset.

Once access has been granted and an appointment made, the researcher will need to “rehearse” the CIT interview. Although the interview is unstructured, there is a sequence of development of the interview that commences with the interviewer introducing himself or herself, his or her research, and the nature of the CIT interview to the subject. Issues such as CONFIDENTIALITY and associated assurances are essential to win the trust of the subject. Once the interviewee is relaxed and comfortable with the procedure, the interview may commence.

The interviewee will then be asked to focus on one or more (usually no more than four) “incidents” that have occurred (over a specified period of time) to the subject or a group or organization of which the subject has intimate knowledge. Different ploys may be used to get the subject to focus on critical events. Respondents vary greatly in the extent to which they can articulate events, and the toughest interviews tend to be those in which there appear to be no events! The interviewer must then attempt to explore lengthy periods of apparent absence of incident to yield useful information. A further difficulty is that of establishing a correct chronology or sequencing of events. This can be facilitated by reference to other events, documents, diaries, and other validating materials. The subject may be encouraged to supply such cross-referencing material.

This helps with TRIANGULATION and VALIDATION of the information.

The respondent may also think of incidents as positive or negative experiences. How subjects frame the incident is entirely under their control, as indeed is the language in which they couch their account. Although it is important that the researcher captures evaluative language, it is also important that he or she listens carefully and PROBES appropriately to grasp the essential details of the incident from the subject's perspective.

Controlling the interview despite its unstructured nature is important. The interview may use generic probes (e.g., What happened next? How did you cope?), which yields information about the context, language, and rapport. When the interviewer does not understand, he or she must interject to get immediate clarification; otherwise, the thread may be lost, and it is often difficult to return to more than a few minor points at the end of an interview. It can also be difficult to remember something that one has not understood! When finally concluding the CIT, the interviewer must take care of all ethical issues and concerns. If ANONYMITY is important, then assurances must be given. If feedback is required in some form, then an arrangement must be made to ensure that it is given. The interviewer must leave the interview feeling that, at whatever time or stage in the future, he or she will be welcome to return.

The final stage in the process is TRANSCRIPTION of the interview and its analysis. The analytical process is likely to be based on GROUNDED THEORY. This means that the researcher abandons preconceptions and, through the process of analysis, builds up an explanatory framework through the CONCEPTUALIZATION of the data. This is THEORY building in its most basic form. It should be used ideally with THEORETICAL SAMPLING techniques so that depth of understanding can be developed from which a coherent body of theory will eventually be formed. The CIT is not easy to conduct well and requires the researcher to be skilled and mature in approach. It can be stressful, but on the whole, it is enjoyable both for the researcher and subject. It is a tool that helps us develop an understanding of dynamic processes within social intercourse. It assumes that real life is messy and that all, including the researcher and the subject, are attempting to make sense of their reality.

—Elizabeth Chell

REFERENCES

- Andersson, B. E., & Nilsson, S. G. (1964). Studies in the reliability and validity of the critical incident technique. *Journal of Applied Psychology*, 48(1), 398–403.
- Chell, E. (1998). Critical incident technique. In G. Symon & C. Cassell (Eds.), *Qualitative methods and analysis in organizational research: A practical guide* (pp. 51–72). London: Sage.
- Chell, E., & Baines, S. (1998). Does gender affect business performance? A study of micro-businesses in business services in the U.K. *International Journal of Entrepreneurship and Regional Development*, 10(4), 117–135.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, 51(4), 327–358.
- Ronan, W. W., & Latham, G. P. (1974). The reliability and validity of the critical incident technique: A closer look. *Studies in Personnel Psychology*, 6(1), 33–64.

CRITICAL PRAGMATISM

Critical pragmatism is a methodological orientation that believes that social science research should illuminate ideological domination, hegemonic practices, and social injustice; it also advocates an eclectic methodological experimentalism in the pursuit of this illumination.

The critical element in critical pragmatism is derived from CRITICAL THEORY, particularly the work of Herbert Marcuse and Antonio Gramsci. Both these writers urge the need for research into the ways in which societies reproduce themselves and the ways in which people are persuaded to embrace ideas, beliefs, and practices that are patently not in their best interests. Marcuse was interested in the way in which people could learn to challenge the widespread existence of one-dimensional thought—that is, thought that prevented any consideration of the moral basis of society and that dissuaded people from utopian speculation (see Marcuse, 1964). Gramsci was interested in the way people learned to recognize and contest hegemony—that is, the process by which people embrace ways of thinking and acting that are harmful to them and that keep unequal structures intact (see Gramsci, 1971). The pragmatic element comes from the tradition of American PRAGMATISM. Pragmatism has an experimental, anti-foundational tenor that advocates using an eclectic range of methods and approaches to promote democracy. It encourages the constant critical analysis of assumptions and is skeptical of reified, standardized models of practice or inquiry.

The element of FALLIBILISTIC self-criticality endemic to critical pragmatism means avoiding a slavish adherence to a particular methodology, whether this be QUANTITATIVE, QUALITATIVE, POSITIVIST, ETHNOGRAPHIC, statistical, or PHENOMENOLOGICAL. A critically pragmatic stance rejects any premature commitment to one research approach, no matter how liberatory this might appear. It urges a principled methodological eclecticism, a readiness to experiment with any and all approaches in the attempt to illuminate and contest dominant ideology.

Critical pragmatism has as the focus of its research effort a number of questions. How do people learn forms of reasoning that challenge dominant ideology, and how do they learn to question the social, cultural, and political forms that ideology justifies? How do people learn to interpret their experiences in ways that emphasize their connectedness to others and lead them to see the need for solidarity and collective organization? How do people learn to unmask the flow of power in their lives and communities? How do people learn of the existence of hegemony and of their own complicity in its continued existence? Once aware of hegemony, how do they contest its all-pervasive effects? A quantitatively inclined example of critical pragmatism would be John Gaventa's (1980) study of land ownership in Appalachia. A qualitatively inclined example would be Douglas Foley's (1990) ethnography of Mexican American youth in a south Texas town.

—Stephen D. Brookfield

REFERENCES

- Foley, D. E. (1990). *Learning capitalist culture: Deep in the heart of Texas*. Philadelphia: University of Pennsylvania Press.
- Gaventa, J. (1980). *Power and powerlessness: Quiescence and rebellion in an Appalachian valley*. Chicago: University of Chicago Press.
- Gramsci, A. (1971). *Selections from the prison notebooks*. London: Lawrence and Wishart.
- Marcuse, H. (1964). *One dimensional man*. Boston: Beacon.

CRITICAL RACE THEORY

Critical race theory is a body of radical critique against the implicit acceptance of White supremacy in prevailing legal paradigms and in contemporary law. The legal literature and civil rights

discourse used to be dominated by the writings of White scholars and legal practitioners. Critical race theory emerged in the early 1990s as an intellectual movement shaped substantially, but not exclusively, by progressive scholars of color in the United States (Crenshaw, Gotanda, Peller, & Thomas, 1995). Critical race theory deconstructs the legal, symbolic, and material capital represented by *Whiteness*. Its mission is to critique and, ultimately, transform the law, an institution contributing to the construction and perpetuation of racial domination and subordination. The intertwining of experiential depth, politics, and intellectual rigor makes critical race theory accessible and relevant to social scientists despite its foundations in the discipline of law.

Critical race theory can be seen as part of the radical tradition of oppositional writings (Essed & Goldberg, 2002). These approaches are premised on the idea that science is not “neutral” or “objective.” Notions of rights and equality, right and wrong, justice and injustice have different implications for groups who have suffered through history. An important methodological consideration is that race and racism are often understood differently, depending on the positioning of the defining party in the matrix of race relations (Bulmer & Solomos, 2003). Institutions, the court included, tend to privilege the perpetrator's perspective of racism, usually a narrow definition based on proof of intent. The victim, who is exposed to a broader range of experiences of racism, can often understand racism in terms of the societal consequences of exclusion and humiliation.

Critical race theory has been instrumental in mainstreaming *intersectionality* and *multiple identifications* as theoretical frameworks. Race, gender, and other structural categories intersect with and determine legal and social discourses. For example, the disproportionate number of African American women who lose custody of their children through the child welfare system is partly due to cultural bias privileging the White and middle-class ideal of the nuclear family and (full-time) mothering.

Using texts and narratives as data, critical race scholars expose the limitations and distortions of traditional legal analysis. They deconstruct law through contextual analysis of historical and other court cases that have marked the landscape of race relations. The myth of *color blindness* underlying the well-known “separate but equal” case, *Plessy v. Ferguson* (1896), is a case in point. The court treated racial

categorizations as unconnected to social status and reconfirmed segregation. Critical race theory offers a framework for understanding *how* and *why* court systems continue to reinforce racial injustice. Collection and analysis of historical and contemporary texts and stories of racial injustice are powerful tools to build bridges of validation and understanding in the pursuit of social transformations.

—Philomena Essed

REFERENCES

- Bulmer, M., & Solomos, J. (Eds.). (2003). *Researching race and racism*. London: Routledge Kegan Paul.
- Crenshaw, K., Gotanda, N., Peller, G., & Thomas, K. (Eds.). (1995). *Critical race theory*. New York: The New York Press.
- Essed, Ph., & Goldberg, T. D. (Eds.). (2002). *Race critical theories: Text and context*. Malden, MA: Blackwell.
- Plessy v. Ferguson*, 163 U.S. 537 (1896).

CRITICAL REALISM

Critical realism is a way of understanding the nature of both natural and human social sciences. Particularly in the case of the latter, it also advocates some and argues against other ways of practicing them. As a version of REALISM, it is committed to the view that the objects of scientific knowledge both exist and act independently of our beliefs about them. However, the qualification *critical* serves to distinguish this form of realism from others, sometimes called *direct* or *naive* realisms, which assume simple “one-to-one” links between beliefs and reality. In the human social sciences, the qualifier *critical* also often signals a normatively critical orientation to existing social forms.

Critical realists make a clear distinction between the independently existing real beings, relations, processes, and so on (the “intransitive dimension”), which are the objects of scientific knowledge, and the socioculturally produced concepts, knowledge claims, and methods through which we attempt to understand them. The latter (sometimes referred to as the *transitive dimension*) (Bhaskar, 1979/1998) are always provisional and fallible. Critical realists thus fully acknowledge the socially and historically “constructed” character of scientific theoretical frameworks and knowledge claims, but they argue that practices such as experiment, scientific education,

and the application of scientific ideas presuppose the independent *existence* of the objects of scientific knowledge. We cannot make sense of these practices other than on the basis that they are *about* something, even though we may doubt the truth of specific assertions about that “something.”

In most cases, critical realists reject the relativism often associated with approaches (such as their own), which emphasize the socially constructed character of knowledge. Without commitment to unsustainable notions of “ultimate” or “absolute” truth, it is usually possible to evaluate rival knowledge claims as (provisionally) more or less rationally defensible in light of existing evidence.

Critical realists have provided powerful arguments against both empiricist and relativist accounts of the nature of natural scientific activity. Bhaskar (1975/1997), in particular, used transcendental arguments from practices such as scientific experiment to show the incoherence of empiricist restrictions on scientific ONTOLOGY and accounts of scientific laws. Experimental practice is designed to isolate causal mechanisms, so producing regular event sequences, or “constant conjunctions.” But because these constant conjunctions are the outcome of experimental manipulation, they cannot be identical with scientific laws (otherwise, we would have to say that laws of nature were products of human manipulation).

This analysis results in a view of the world as “stratified,” distinguishing between the “level” of observed events (the “empirical”) and that of the “actual” flow of events (most of which pass unobserved). A third level, misleadingly referred to as the “real” (because all levels are in some sense real), is that of the “causal mechanisms,” whose tendencies are captured by statements of scientific laws. Critical realism might reasonably be described as a form of “depth” realism, implying the existence of multiple layers of reality behind or below the flow of sense-experience and subject to discovery by scientific experiment and theoretical analysis. Science is, in this way, seen as a process of discovery of ever deeper levels of reality (genes, molecules, atoms, subatomic particles, fields of force, quasars, etc.) rather than as one of accumulating experiential generalizations. The analysis of practices of applying scientific knowledge in technology yields a further claim about the nature of the world as an object of scientific knowledge. This is its “differentiated” character. Experimental practice involves isolating causal mechanisms from extraneous influences (i.e., the creation of “closed

systems”). However, technologies involve application of scientific knowledge in “open systems,” involving the interaction of multiple causal mechanisms. In open systems, interaction between mechanisms results in the overriding or modification of their effects, such that regular or predictable flows of events and observations may not occur.

The applicability of the critical realist model of science to the social sciences is contested. For some philosophers of social science, humans and their social relations are so unlike the natural world that it is fundamentally inappropriate to attempt to study them on the basis of any model of science. However, critical realists have generally taken the view that despite important differences between their objects of knowledge and those of the natural sciences, the social studies may still be scientific. However, if we recognize that each science must adopt methods of enquiry and forms of explanation appropriate to its subject matter, it still remains an open question just what it is to study social life in a (critical realist) scientific way.

During the 1970s, several authors addressed this question from realist points of view (e.g., Benton, 1977; Bhaskar, 1979/1998; Keat & Urry, 1975). Bhaskar argued that the social sciences can be scientific—not just *despite* but also *because* of acknowledged differences between nature and society, as well as our relationship to them. Social structures are held to exist only in virtue of the activity of agents, are dependent on agents’ beliefs about them, and are only relatively enduring. There are also problems about the partial identity of the observer and what is observed: Social sciences are themselves part of society, and it is not possible to establish experimental closure in the social sciences.

These differences, for Bhaskar (1979/1998), imply only that the social sciences must take account of them in its methodology—for example, by beginning investigations with actors’ own understanding of what they are doing. For him, social crises constitute an analog of experimentation in the natural sciences. Collier (1994), however, regards the lack of a social science counterpart to experimentation in the natural sciences as grounds for denying that they (the social “sciences”) can be fully scientific: He terms them *epistemoids*. Outhwaite (1987) argues for a close relationship between critical realism and critical theory, whereas Benton (1998) sees in critical realism a basis for a nonreductive realignment of the social and natural sciences, which is important for understanding ecological issues in particular.

Another area of disagreement among critical realists concerns the relationship between explanation and moral or political values. Bhaskar and others argue that it is possible to derive values from scientific explanation on the realist model—hence the possibility of “critical” social science, which is oriented to a vision of human emancipation. Marx’s “explanatory critique” of ideology is the best-known example: Other things being equal, social or economic relations that tend to produce false beliefs in their agents are undesirable and ought to be transformed. Against this, Sayer (2000) has argued that existing social forms are only legitimately open to criticism on the basis of accounts of plausible alternatives to them.

In general, for those social scientists who espouse critical realism, there is a strong commitment to the distinction between social structures and human agency. Both are regarded as real, distinct, but interdependent levels. Agents both reproduce and transform social structures through their activities, whereas social structures both enable and constrain social action (Archer, 1995). There is a large and growing body of empirical research in geography, economics, law, sociology, politics, psychology, feminist and environmental studies, and other fields informed by critical realism (for examples of the application of critical realism, see New, 1996; Porter, 1993).

—Ted Benton

REFERENCES

- Archer, M. (1995). *Realist social theory: The morphogenetic approach*. Cambridge, UK: Cambridge University Press.
- Archer, M., Bhaskar, R., Collier, A., Lawson, T., & Norrie, A. (Eds.). (1998). *Critical realism: Essential readings*. London: Routledge Kegan Paul.
- Benton, T. (1977). *Philosophical foundations of the three sociologies*. London: Routledge Kegan Paul.
- Benton, T. (1998). Realism and social science. In M. Archer, R. Bhaskar, A. Collier, T. Lawson, & A. Norrie (Eds.), *Critical realism: Essential readings*. London: Routledge Kegan Paul.
- Bhaskar, R. (1997). *A realist theory of science*. London: Verso. (Original work published 1975)
- Bhaskar, R. (1998). *The possibility of naturalism*. Hemel Hempstead, UK: Harvester Wheatsheaf. (Original work published 1979)
- Collier, A. (1994). *Critical realism*. London: Verso.
- Keat, R., & Urry, J. (1975). *Social theory as science*. London: Routledge Kegan Paul.
- New, C. (1996). *Agency, health and social survival*. London: Taylor & Francis.

- Outhwaite, W. (1987). *New philosophies of social science*. London: Macmillan.
- Porter, S. (1993). Critical realist ethnography: The case of racism and professionalism in a medical setting. *Sociology*, 27, 591–609.
- Sayer, A. (2000). *Realism and social science*. London: Sage.

CRITICAL THEORY

To the extent that any theorization of societal mechanisms and modes of conducting social relations does not accord with dominant ways of viewing society and social relations, it may be said to be critical. The term *critical theory*, however, is normally reserved for a particular set of ideas associated with what become known as the Frankfurt School of Social Research and its followers, who have modified and extended its original insights, aspirations, and agendas.

Under the auspices of its second director, Max Horkheimer (1895–1973), the Institute for Social Research at Frankfurt centered its interests on the following key areas: EXPLANATIONS for the absence of a unified working-class movement in Europe, an examination of the nature and consequences of capitalist crises, a consideration of the relationship between the political and the economic spheres in modern societies, an account for the rise of fascism and Nazism as political movements, the study of familial socialization, and a sustained critique of the link between POSITIVISM and science. To pursue this study, those figures who stood between Hegel, Marx, and the Frankfurt School—such as Schopenhauer and Nietzsche, who had both questioned the Enlightenment project—required systematic engagement to recover the promise of Marxism.

As noted, the changing climate shaped the work of these scholars. In Nazi Germany, the institute found itself under threat with the result that its leading members, including Theodor Adorno (1903–1969), Max Horkheimer, and Herbert Marcuse (1898–1979), emigrated to the United States (Adorno and Horkheimer returned after World War II). Here they found a self-confident bourgeois liberal-capitalism, with its apparent ability to absorb and neutralize proletarian consciousness. Here was the personification of the ideology of individualism accompanied by ideas of success and meritocracy. Analyzing these trends and their consequences necessitated a consideration of culture that, until this time, had been largely devalued in Marxist circles.

The work of these scholars was interdisciplinary (thus anticipating debates that are more prominent in contemporary times), accompanied by critiques of both positivist and interpretivist approaches to understanding human relations. On one hand, positivism had failed to examine the conditions under which capitalism develops and is sustained. On the other hand, interpretivism was held to be inadequate due to an exclusive concentration on the process of self-understanding and self-consciousness. The result was an uncritical acceptance of dominant forms of consciousness within given societies and a failure to consider the structural determinants of human actions.

Historical investigations were then undertaken into the Enlightenment project. Here we find that

the individual is wholly devalued in relation to the economic powers, which at the same time press the control of society over nature to hitherto unsuspected heights. . . . The flood of detailed information and candy-floss entertainment simultaneously instructs and stultifies mankind. (Adorno & Horkheimer, 1944/1979, pp. xiv–xv)

On this basis, studies examined the process of ideological incorporation. Herbert Marcuse (1968), in *One Dimensional Man*, thus spoke of the “technical-administrative control” of society.

This meta-critique noted the following question: Having rejected interpretivist social theory, how can the important analysis of mental states be achieved? This is where Sigmund Freud enters as part of the whole history of the relationship between psychoanalysis and social theory. This focus moves away from an understanding of the solitary ego in the social world to the *self-misunderstanding* person who fails to see the causes of his or her own symptoms. However, issues remained. Up to the time of the Frankfurt theorists, Marxists had dismissed psychoanalysis as unworthy of attention and a distraction from the primary purpose of overthrowing an unjust economic system. Psychoanalysis, after all, may be seen as the practice of amelioration, not resolution, let alone revolution. Conceptually speaking, it starts from the individual and tends to play down social conditions and constraints.

Therefore, if Freud’s work was to serve the needs of critical theory, it required modification. At this point, Leo Lowenthal and Eric Fromm enter the scene as central to this process in the development of critical

theory. Fromm, for example, examined the relationship between the individual and totalitarianism. He also held throughout his life that our capacity for love and freedom is inextricably bound up with socioeconomic conditions. The interest in Freud was also apparent in sociopsychological studies on the structure of modern personality types carried about by Adorno and his associates in 1950, as well as in the writings of Willhelm Reich on the relationship between sexuality and capitalism.

For critical theory, there is a constant interaction between THEORY and facts, and theorists seek to recognize the relationship between the constitution of their propositions and the social context in which they find themselves. REFLEXIVITY concerning the relationship between what is produced, under what circumstances, and with what effects distinguishes traditional from critical theory. In addition, the issue of research results feeding back into social life is not a “problem” for researchers. On the contrary, the adequacy of critical research lies in its value for informing political actions.

Although critical social research aligns itself with the “wishes and struggles of the age,” it remains the case that critics, although recognizing the impossibility of separating values from research, regard justifications for conflating the two as untenable. After all, who is to define the values that are of importance? Furthermore, what is the basis of critique? Is it the autonomy of the free individual? The idea of autonomy becomes an abstraction because how can any individual be presumed to be separate from the social relations of whom he or she is a part? The basis of critique then becomes so far removed from reality that it appears aloof and withdrawn. These and other criticisms have been made of research practice informed by critical theory.

Of all contemporary scholars, Jürgen Habermas has continued most prolifically in this tradition and addressed the above issues through extensive modification and critique. This legacy is also evident in the writings of Nancy Fraser (1997) and Axel Honneth (1996), who have contributed to debates on recognition and redistribution, as well as Pierre Bourdieu's (2000) contributions to a “realpolitik of reason.” As for Michel Foucault, his own words in “Remarks on Marx” will suffice: “The Frankfurt School set problems that are still being worked on. Among others, the effects of power that are connected to a rationality that has been historically and geographically defined in the

West, starting from the sixteenth century on” (Foucault, 1991, p. 117).

—Tim May

REFERENCES

- Adorno, T., & Horkheimer, M. (1979). *Dialectic of enlightenment* (J. Cumming, Trans.). London: Verso. (Original work published 1944.)
- Bourdieu, P. (2000). *Pascalian meditations* (R. Nice, Trans.). Cambridge, UK: Polity.
- Comstock, D. E. (1994). A method for critical research. In M. Martin & L. C. McIntyre (Eds.), *Readings in the philosophy of social science* (pp. 625–639). Cambridge: MIT Press.
- Foucault, M. (1991). *Remarks on Marx: Conversations with Duccio Trombadori* (R. J. Goldstein & J. Cascaito, Trans.). New York: Semiotext(e).
- Fraser, N. (1997). *Justice interruptus: Critical reflections on the ‘postsocialist’ condition*. London: Routledge Kegan Paul.
- Honneth, A. (1996). *The struggle for recognition: The moral grammar of social conflicts* (J. Anderson, Trans.). Cambridge: MIT Press.
- Marcuse, H. (1968). *One dimensional man: The ideology of industrial society*. London: Sphere.
- Morrow, R. A. (with Brown, D. D.). (1994). *Critical theory and methodology*. London: Sage.
- Wiggershaus, R. (1995). *The Frankfurt School: Its history, theories and political significance* (M. Robertson, Trans.). Cambridge, UK: Polity.

CRITICAL VALUES

In a HYPOTHESIS TEST, we decide whether the data are extreme enough to cast doubt on the NULL HYPOTHESIS. If the data are extreme, we reject the null hypothesis; if the data are moderate, we accept it. A threshold between extreme and moderate data, between accepting and rejecting the null hypothesis, is called a *critical value*.

Suppose we flip a coin 10 times. An obvious null hypothesis is that the coin is fair, or equally likely to come up heads or tails. If the null hypothesis is true, we will probably flip close to 5 heads. A sensible test would accept the null hypothesis if the number of heads is close to 5 (moderate) and reject the null hypothesis if the number of heads is far from 5 (extreme). For example, a test might reject the null hypothesis if we flip 9 heads or more or if we flip 1 head or fewer. For this test, 1 and 9 are the critical values.

Critical values are related to SIGNIFICANCE LEVELS. The test above has a significance level of .0215: Under

the null hypothesis, there is a .0215 probability that the number of heads will be at least as extreme as the critical values. (This probability is calculated using the BINOMIAL DISTRIBUTION.) We can use the critical values to calculate the significance level, as we did here. It is more common, however, to choose the significance level first and then calculate the corresponding critical values.

Tests using critical values are equivalent to tests using *p* VALUES. That is, rejecting a hypothesis because of extreme data is equivalent to rejecting the hypothesis because of a small *p* value. For example, suppose 10 coin flips produced 10 heads; 10 is more extreme than the critical value of 9, so we would reject the null hypothesis using the critical value. But we would also reject the null hypothesis if we used the TWO-TAILED *p* value: The *p* value for 10 heads is .002, which is less than the significance level of .0215. As this example illustrates, decisions based on *p* values will always agree with decisions based on critical values.

—Paul T. von Hippel

CRONBACH'S ALPHA

Cronbach's alpha is one type of INTERNAL RELIABILITY estimate used to assess the consistency of responses on a composite measure that contains more than one component (i.e., items, raters). Alpha coefficient, the most widely used reliability estimate, was first introduced by the late Lee J. Cronbach in 1951. According to the Social Sciences Citation Index, between 1951 and 2002, the article was cited 4,912 times by other published articles. The number of published and unpublished articles that report Cronbach's alpha without citing the 1951 article may very likely approach an astronomical figure.

Cronbach's alpha and standardized Cronbach's alpha are formulated as

$$\alpha = \left(\frac{k}{k-1} \right) \left(1 - \frac{\sum_{i=1}^k s_i^2}{s_t^2} \right)$$

and standardized

$$\alpha = \frac{k\bar{r}}{1 + \bar{r}(k-1)},$$

where *k* is the number of components in a composite measure, s_i^2 is the variance of component *i*, s_t^2 is the

variance of the total score on the measure, and $\bar{r}$ is the average intercomponent correlation. Standardized Cronbach's alpha is used when component standardized scores are computed instead of using raw component scores. The precision of Cronbach's alpha is determined by the standard error of the intercomponent correlations, formulated as

$$\frac{s_r}{\sqrt{[0.5 \times k \times (k-1)] - 1}},$$

where s_r is the standard deviation of the intercomponent correlations. All other things being equal, Cronbach's alpha becomes less precise if the variation among intercomponent correlations increases. It should be noted, however, that the size of Cronbach's alpha is not a function of the size of the standard error. It is possible to have two sets of *k* components with different amounts of variation among their intercomponent correlations but similar Cronbach's alpha coefficients, which are determined by the average intercomponent correlation.

Conceptually, Cronbach's alpha, rather than standardized Cronbach's alpha, is a lower bound estimate of the *proportion of variance* in a composite measure that is accounted for by a common factor underlying all components. Although this statement suggests that Cronbach's alpha should not be negative and can only vary between 0 and 1, it is not unlikely in practice to observe a negative Cronbach's alpha. As shown in Table 1, a seven-item measure is administered to a group of 10 people. Each item has three response categories, ranging from one to three. The alpha coefficient is negative, and its absolute value is even greater than 1.

Cronbach's alpha can be affected by characteristics of respondents and components. Everything being equal, if respondents are homogeneous (i.e., similar in the characteristic assessed by a measure), s_i^2 tends to decrease. As seen from the above formulas, a small s_i^2 tends to lead to a small alpha coefficient. Second, Cronbach's alpha tends to increase as the number of good-quality components increases. However, this positive feature becomes less salient when the number of components is so large that $\frac{k}{k-1}$ approaches unity. Note that if the components on a measure are heterogeneous, suggesting there is more than one factor or dimension underlying the components, Cronbach's alpha will underestimate the true reliability. In addition, a META-ANALYTIC finding by Peterson (1994) suggests that Cronbach's alpha may underestimate the true reliability when the measure is composed of less

Table 1 A Demonstration of a Negative Cronbach's Alpha

Person	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Total
A	1.00	2.00	2.00	2.00	1.00	2.00	2.00	12
B	2.00	1.00	2.00	3.00	2.00	3.00	2.00	15
C	1.00	2.00	1.00	2.00	2.00	2.00	2.00	12
D	2.00	3.00	2.00	1.00	2.00	1.00	2.00	13
F	1.00	2.00	3.00	2.00	1.00	2.00	3.00	14
G	3.00	1.00	2.00	3.00	2.00	1.00	2.00	14
H	1.00	2.00	2.00	2.00	2.00	2.00	1.00	12
I	2.00	1.00	2.00	1.00	2.00	1.00	2.00	11
J	1.00	2.00	1.00	2.00	3.00	2.00	1.00	12
K	1.00	3.00	2.00	1.00	2.00	2.00	2.00	13
Variance	0.45	0.49	0.29	0.49	0.29	0.36	0.29	1.36

$$\alpha = \left(\frac{k}{k-1} \right) \left(1 - \frac{\sum_{i=1}^k s_i^2}{s_t^2} \right) = \left(\frac{7}{6-1} \right) \left(1 - \frac{2.66}{1.36} \right) = -1.12$$

$$\text{Standardized } \alpha = \frac{k\bar{r}}{1 + \bar{r}(k-1)} = \frac{7 \times (-.0802)}{1 + (-.0802)(7-1)} = -1.08$$

than four items with only two response categories. It should be emphasized that a high Cronbach's alpha (e.g., 0.8) does not necessarily suggest that there is only one factor or dimension underlying the components. High internal consistency is just a necessary but not sufficient condition for unidimensionality. Another related but often-neglected concern is that Cronbach's alpha is not an appropriate reliability estimate for a composite measure when its components are causal indicators of a CONSTRUCT. For instance, a list of behaviors such as watch TV, read book, ski, visit friends, and play chess describes the construct of social activities. However, these activities cannot be considered as parallel components. Therefore, Cronbach's alpha will not provide an adequate reliability estimate for this measure.

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, 5, 343–355.
- Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, 21, 381–391.

CROSS-CULTURAL RESEARCH

Cross-cultural research can be defined in an anthropological sense to mean any kind of description or comparison of different cultures. It can also be used in the sense of systematic comparisons that explicitly aim to answer questions about the incidence, distribution, and causes of cultural variation (Ember & Ember, 2001). Although it can be useful to identify recurrent themes and patterns in cultures, it can also be risky to generalize about cultural similarities because of the danger of stereotyping. Patterns can provide overall structure to cross-cultural understandings, but they can also hide great individual variations within each culture.

Whether the goal is to describe a unique culture that is different from that of the researcher or to make comparisons across cultures, many issues arise in the conduct of cross-cultural research. Ethnocentrism, or using one's own culture as a lens or standard to judge others, is a barrier to understanding another's cultural experience. Cultural relativism stresses the variability of culture and emphasizes the need to try to understand people in other cultures from within their own cultural context.

Variability within samples in cross-national comparisons, as well as variability on dimensions other than culture, can threaten the VALIDITY of comparative studies (Hujala, 1998). In a study of teachers' thinking

between the United States and Finland, the researchers wanted to use the national political borders to define culture as their independent variable. However, the teachers in Finland and the United States differed on other important dimensions. The Finnish teachers had more training compared to the U.S. teachers, and this difference could have explained differences in the teachers' thinking rather than the cultural thinking in a certain society. Gopaul-McNicol and Armour-Thomas (2002) suggest that researchers examine the relationships between culture and class, culture and ethnicity, and culture and language to better understand how to observe meaningful differences within and between cultural groups. They raise issues related to dominant and subdominant cultures within a country and the EPISTEMOLOGICAL assumptions associated with each group. Development of research methods and data collection instruments by the dominant cultural group can result in a relative neglect of individuals' experiences in the nondominant group.

A debate in the cross-cultural research community reflects contrasting positions concerning the extent to which cross-cultural measures can be developed (Evans, 2000). One position is that tests can be transported from one culture to another with appropriate linguistic translation, administration by a native tester, and provision of familiar content. Cultural psychologists adopt an alternative position because they emphasize the importance of symbolic culture, that is, shared value and meaning, knowing, and communication (Greenfield, 1997). Thus, a test can be transported to another culture only if there is universal agreement on the value or merit of particular responses to particular questions, and the same items must mean the same things in different cultures, given a good linguistic translation of the instrument. In addition, the definition of relevant information must be universally the same, and communication with strangers must be acceptable.

To obtain valid data in cross-cultural research, one must determine the local definitions of concepts (Evans, 2000). The measurement of intelligence in Western countries and in rural Kenyan villages serves as one example of the importance of determining local definitions of concepts. Villagers in rural Kenya describe children's intelligence in terms of their ability to identify animals and trees by characteristics, use, and names and by how far they can carry things, such as stools or food, for a specified distance. These are the criteria that villagers use to determine if a child's growth is on track at the age for school entry.

Cultural representation in the research project is important to the conduct of credible cross-cultural research. This principle applies to studies that are conducted within countries where a variety of cultures are present, as well as to cross-national studies. This should entail full participation in the study from the beginning of the process. One example of a cross-cultural study that followed this principle is the ongoing Preprimary Study of the International Association for the Evaluation of Education Achievement (Olmsted & Weikart, 1994). The study is designed to investigate the nature and effects of various kinds of early childhood experiences on the long-term development of young children. The process and instruments for the study were developed by a cross-cultural team of researchers representing each of the countries initially included in the study (i.e., Belgium, China, Finland, Germany, Hong Kong, Italy, Nigeria, Portugal, Spain, Thailand, and the United States). The group decided on the items to include, as well as the examples and wording that would best represent their culture.

In cross-cultural research, validity in measurement can be influenced by the translation of the instrument or by the influence of cultural norms in behaving with a stranger or an elder (Evans, 2000). A linguistically accurate translation may still lead to misunderstanding because of the multiple meanings attached to words. The translation should include a translation not only of the words but also of the meaning. Work with children may especially be influenced by cultural norms if adults do not generally solicit the views of children or if children are not normally expected to speak in the presence of adults. The researcher in such a context might consider a performance assessment that does not involve asking direct questions.

—Donna M. Mertens

REFERENCES

- Ember, C. R., & Ember, M. (2001). *Cross-cultural research methods*. Walnut Creek, CA: AltaMira.
- Evans, J. L. (2000). *Early childhood counts*. Washington, DC: World Bank.
- Gopaul-McNicol, S.-A., & Armour-Thomas, E. (2002). *Assessment and culture: Psychological tests with minority populations*. San Diego: Academic Press.
- Greenfield, P. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist*, 52(10), 1115-1124.
- Hujala, E. (1998). Problems and challenges in cross-cultural research. *Acta Universitatis Ouluensis*, 35, 19-31.

Olmsted, P., & Weikart, D. P. (Eds.). (1994). *Families speak: Early childhood care and education in 11 countries*. Ypsilanti, MI: High/Scope Press.

CROSS-LAGGED

Cross-lagged models are widely used in the analysis of PANEL DATA, or data collected more than once on the same individuals or units over time, to provide evidence regarding the direction of causality between variables X and Y and to estimate the strength of the CAUSAL effects of each variable on the other. Cross-lagged models involve the estimation and comparison of CORRELATION and/or REGRESSION COEFFICIENTS between each variable measured at one panel wave and the other variable at the *next* panel wave. The underlying logic of such models is that, if X is a cause of Y , then X measured at Time 1 should be related to Y at Time 2; that is, the cross-lagged relationship between X_1 and Y_2 should be nonzero. Furthermore, if X is a more powerful cause of Y than Y is of X , then the cross-lagged relationship between X_1 and Y_2 should be stronger than that between Y_1 and X_2 .

A cross-lagged model for a two-wave panel is shown in diagram form in Figure 1. Variables X_2 and Y_2 from Wave 2 are each hypothesized to be determined by their Wave 1 values, the LAGGED value of the other variable, and an ERROR term U ; the Wave 1 correlation between variables X and Y is represented by $\rho_{X_1Y_1}$, and the correlation between the equations' error terms is represented by $\rho_{U_1U_2}$. The two equations may be written as follows:

$$Y_2 = \beta_1 X_1 + \beta_2 Y_1 + U_1, \quad (1a)$$

$$X_2 = \beta_3 Y_1 + \beta_4 X_1 + U_2, \quad (1b)$$

with the intercepts omitted if the variables are expressed in mean deviation form.

It was originally argued that the causal direction between the two variables could be determined by comparing the magnitude of the two cross-lagged *correlations* in the model, that is, whether $\rho_{X_1Y_2}$ is larger or smaller than $\rho_{Y_1X_2}$. Numerous scholars, however, pointed out that the cross-lagged correlations are a function of the causal effects β_1 (or β_3) in Figure 1, as well as the *stability* of each variable over time (i.e., $\rho_{X_1X_2}$ or $\rho_{Y_1Y_2}$) and the initial correlation between X and Y ($\rho_{X_1Y_1}$). Consequently, REGRESSION and STRUCTURAL EQUATION MODELS are the more

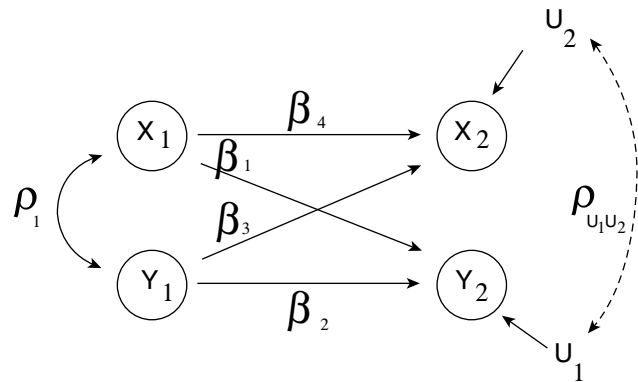


Figure 1 A Two-Wave Panel Model With Cross-Lagged Effects

widely used methods for causal direction and strength, although there are still uses for cross-lagged correlational analysis in testing certain models of spurious association between X and Y (Kenny, 1979).

There are several important threats to causal inference in the cross-lagged panel model. MEASUREMENT ERROR in Y_1 and X_1 will tend to depress the estimates of the stability of Y and X over time, leading to BIAS in the estimates of the cross-lagged effects as well (Kessler & Greenberg, 1981). Second, as is often the case in LONGITUDINAL data, the error terms U_1 or U_2 may be AUTOCORRELATED, leading to a nonzero relationship between Y_1 and U_1 and/or X_1 and U_2 . More fundamentally, the cross-lagged model may yield erroneous results because the true causal lag between X and Y is much shorter than the time period between measurements (Plewis, 1985). In such cases, it would be appropriate to specify a "SYNCHRONOUS" or "cotemporal" effects model, in which X_2 and Y_2 are hypothesized to influence each other in the *same* time period. Models with both cross-lagged and synchronous effects may also be specified and estimated either with INSTRUMENTAL VARIABLES or related procedures or, in the multiwave case, by imposing equality or other constraints on the parameters across waves.

—Steven E. Finkel

REFERENCES

- Kenny, D. A. (1979). *Correlation and causality*. New York: John Wiley.
- Kessler, R., & Greenberg, D. (1981). *Linear panel analysis*. New York: Academic Press.

Plewis, I. (1985). *Analysing change: Measurement and exploration using longitudinal data*. Chichester, UK: Wiley.

CROSS-SECTIONAL DATA

Cross-sectional data come from a study in which all observations for each participant are collected at approximately the same point in time and in which time is not an element in the RESEARCH DESIGN. All observations are considered to have been made concurrently, although literally this is never true. Even within a single session, an individual completes items in a sequence over time, and within a study, all participants do not typically provide data at the same moment. However, the time frame in such a design is typically short, and it is presumed that time of observation during the course of the study had RANDOM effects that produced only error. This is contrasted with LONGITUDINAL designs in which observations are taken at different points in time to see if there were differences in variables due to events occurring over the course of the study.

The term *cross-sectional* is often used to describe QUESTIONNAIRE studies. The simplest is the single-source cross-sectional design in which participants provide all data about themselves, and all items and scales are administered in a single session. The multisource design adds DATA not contained in the participant's questionnaire and might include questionnaires given to other people (e.g., coworker or family member), data provided by observers, or archival data. Another approach could be to divide a long questionnaire into two portions that are administered at different sessions but in which there is no attempt to draw inferences about time.

It is possible to have cross-sectional data without questionnaires and in studies in which the sampling unit is not a person. For example, one might look at relations between different variables in the U.S. census in which all data are considered to have been collected concurrently.

Cross-sectional data can be highly efficient for determining if there are relations between VARIABLES. Their biggest limitation is in not allowing confident CAUSAL conclusions (Shadish, Cook, & Campbell, 2002; Spector, 1994). Without having time built into the design, there is no way to determine the direction of causality between two variables that are related or if there are additional variables that account for the

observed relations. For that, one must have either a longitudinal design or an EXPERIMENT.

—Paul E. Spector

REFERENCES

- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
- Spector, P. E. (1994). Using self-report questionnaires in OB research: A comment on the use of a controversial method. *Journal of Organizational Behavior*, 15, 385–392.

CROSS-SECTIONAL DESIGN

A cross-sectional design refers to collecting DATA that are assumed to have been collected at one point in time. The objective is to get a “snapshot” or picture of a group. The data hypothetically can represent individuals, groups, institutions, behaviors, or some other unit of analysis, and the data can be collected using a range of procedures. In actuality, the term *cross-sectional design* is often used interchangeably with the term SURVEY and most frequently refers to data that are collected through interviews or questionnaires that are administered to individuals who either represent themselves or who represent households, institutions, or other social entities.

Although not mandatory, cross-sectional designs often assume that the group or population under study is heterogeneous in its characteristics—that is, people of many ages, behaviors, and opinions are represented within the study POPULATION. The objective is to properly “represent” the diversity within the group. With this assumption of heterogeneity come assumptions about the need to appropriately represent or infer to some larger population or group from which the study subjects are drawn when findings are analyzed and conclusions are drawn. As a result, the procedures by which study subjects are selected are of concern. If, within the cross-sectional study, data are not collected from the entire population of interest, then it is important to know how those chosen for study were selected. PROBABILITY or SYSTEMATIC SAMPLING strategies generally are preferred. Because data in cross-sectional designs are collected at one point in time, the emphasis in analysis is to examine the FREQUENCY DISTRIBUTIONS of single

VARIABLES and the ASSOCIATIONS between two or more variables. Researchers sometimes distinguish between descriptive and EXPLANATORY cross-sectional studies, with explanatory studies moving beyond description to examine whether conclusions can be drawn about how certain things influence or result in change (e.g., examining whether exposure to disasters increases psychological distress).

Cross-sectional designs can be contrasted with group comparison or CASE CONTROL designs; LONGITUDINAL, prospective, or PANEL designs; and true EXPERIMENTAL or QUASI-EXPERIMENTAL designs. The objective of longitudinal, experimental, and group comparison designs is to maximize the ability to make PREDICTIONS, investigate CAUSAL linkages, and “explain” why or how something happens, whereas cross-sectional designs are better able to describe relationships between variables. These designs (e.g., experimental) are less concerned with EXTERNAL VALIDITY, or the need to generalize to a larger population, and more concerned with INTERNAL VALIDITY, or the need to accurately measure the relationships between variables or constructs.

Sociologists conducted most of the early research on how to design cross-sectional studies. Paul Lazarsfeld is considered one of the major contributors to early research on cross-sectional studies, and Leslie Kish made major contributions to how participants in cross-sectional studies should be selected or sampled (see Kish, 1965; Lazarsfeld, 1958). Lazarsfeld was particularly instrumental in developing procedures by which causal inferences could be made from cross-sectional data before the advent of computers. His work on the ELABORATION MODEL was extended by Morris Rosenberg (1968).

—Linda B. Bourque

REFERENCES

- Aday, L. A. (1996). *Designing and conducting health surveys: A comprehensive guide* (2nd ed.). San Francisco: Jossey-Bass.
- Kendall, P. L., & Lazarsfeld, P. F. (1950). Problems of survey analysis. In R. K. Merton & P. F. Lazarsfeld (Eds.), *Continuities in social research: Studies in the scope and method of “the American soldier”* (pp. 160–176). Glencoe, IL: Free Press.
- Kish, L. (1965). *Survey sampling*. New York: John Wiley.
- Lazarsfeld, P. F. (1958). Evidence and inference in social research. *Daedalus*, 87, 120–121.
- Rosenberg, M. (1968). *The logic of survey analysis*. New York: Basic Books.

CROSS-TABULATION

A cross-tabulation is a table that displays the RELATIONSHIP between two variables. It is also called a cross-tab, or a CONTINGENCY TABLE. The simplest is a 2×2 table, consisting of four CELLS (one for each possible combination of values). Suppose one variable is gender (male = 1, female = 0) and the other variable is vote (yes = 1, no = 0). Gender, the presumed INDEPENDENT VARIABLE, can be placed at the top of the table (as the column variable). Vote, the presumed DEPENDENT VARIABLE, can be placed at the side of the table (as the row variable). The raw frequencies and the percentage frequencies can be placed in each cell; for example, in Cell A, there might be 214 men (55%) who voted. Tables contain more cells as the variables have more categories. Also, the analyst might look at a cross-tab between two variables while controlling on a third variable. For example, one might look at the table for gender and vote within the White subsample and within the non-White subsample.

—Michael S. Lewis-Beck

See also CONTINGENCY TABLE

CUMULATIVE FREQUENCY POLYGON

A cumulative frequency polygon or ogive is a variation on the frequency polygon. Although both are used to describe a relatively large set of quantitative data, the distinction is that cumulative frequency polygons show cumulative frequencies on the y-axis, with frequencies expressed in either absolute (counts) or relative terms (proportions). Cumulative frequencies are useful for knowing the number or the proportion of values that fall above or below a given value.

In the sample cumulative frequency polygon in Figure 1, it is easy to judge the number of males shorter than 68 inches and the number at this height or above. Changing the y-axis values to proportions ($200/1,000 = .20$, $400/1,000 = .40$, etc.) would yield the proportion of males shorter than 68 inches and the proportion at this height and above. The same information cannot be obtained so easily in a frequency polygon because the y-axis yields a measure of the

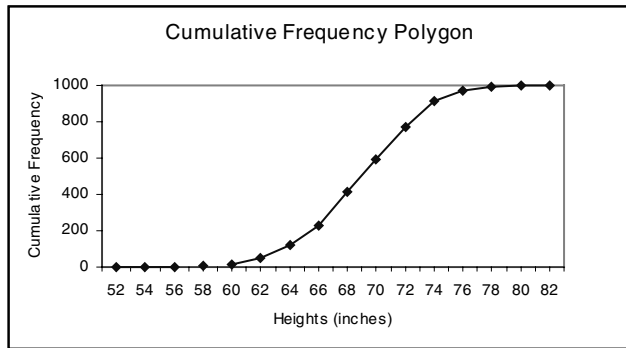


Figure 1 Adult Male Heights Displayed According to the Cumulative Frequency for Each 2-Inch Interval Between 52 and 82 Inches

frequency (absolute or relative) of observations within a single category, not cumulatively across categories.

To create a cumulative frequency polygon, researchers first sort data from high to low and then group them into contiguous intervals. The upper limits to each interval are represented on the x -axis of a graph. The y -axis provides a measure of cumulative frequency. It shows either the number or the proportion of data values falling at or below each upper limit. For absolute frequencies, it is scaled from a minimum of 0 to a maximum that equals the number of measures in the data set. For relative frequencies, it is scaled from 0 to 1.00. Within the field of the figure, straight lines connect points that mark the cumulative frequencies.

Cumulative frequency polygons are used to represent quantitative data when the data intervals do not equal the number of distinct values in the data set. This is always the case for continuous data. When the data are discrete, it is possible to represent every value in the data set on the x -axis. In this case, a cumulative histogram would be used instead of a cumulative frequency polygon. The importance of the distinction is that points are connected with lines in the polygon to allow interpolation of frequencies for x -axis values not at the upper limits (e.g., 67 inches in the figure). When data have no values other than those represented on the x -axis, cumulative frequencies are simply marked by bars (or columns).

Some statisticians reserve the term *cumulative frequency polygon* just for graphs that show the absolute frequencies of values and prefer the term *cumulative relative frequency polygon* for graphs that show proportions. The alternative is to refer, as here, to two types

of cumulative frequency polygons—one for counts and one for proportions.

—Kenneth O. McGraw

CURVILINEARITY

Variables in social science research can exhibit a curvilinear relationship to one another, which means that their relationship changes depending on the numerical values each variable takes. Usually, the change is expressed in terms of values of the INDEPENDENT VARIABLE first and then the impact on the DEPENDENT VARIABLE. These changes in SLOPE appear to take the shape of a curve if plotted on a line, as seen on a two-dimensional graph with the independent variable along the x -axis and the dependent variable along the y -axis. Addressing curvilinear relationships is most critical in ORDINARY LEAST SQUARES (OLS) analysis.

An example of this phenomenon from political science is the relationship between economic development (gross national product [GNP] per capita) and democracy across countries. As very poor countries increase their wealth, they also can dramatically increase their levels of democratic performance. Countries at very low levels of wealth can expect to secure for themselves substantial, linearly steep gains in democratic performance. (Naturally, the relationship is not perfect, as no uncausal explanation is satisfactory in accounting for social science phenomena.) However, the initial dramatic gains in democracy become less dramatic as the country becomes much more economically developed. The graphical effect of this curvilinearity is noticeable. The slope of the relationship shifts from steep at low levels of wealth to nearly zero at high levels of wealth, where most countries cluster around the high values of democracy.

For OLS regression to produce estimates that are best linear unbiased estimates (BLUE), the regression assumptions must be met. One of these assumptions is that the independent variables must have linear relationships with the dependent variable. This assumption is clearly violated in the instance of GNP per capita's relationship to democracy. The resulting OLS estimation will produce biased estimates.

The analyst can pursue remedies to address the problem of curvilinearity in OLS regression, all of which are oriented around arithmetically transforming

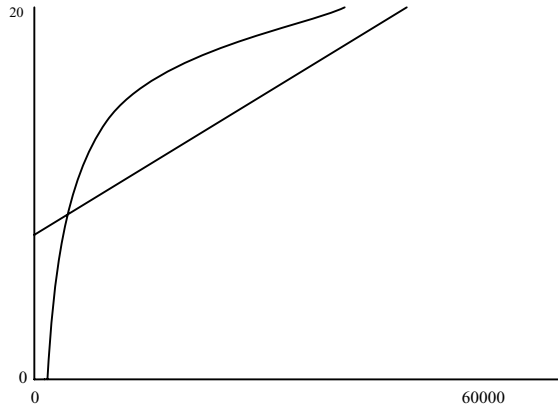


Figure 1 Curvilinear (Versus Linear) Relationship of Economic Development to Democracy

the variable so that the linearity assumption can be met. Social science analysts most commonly employ one of three transformations. The most frequently used is the logarithmic transformation, which can be either to the natural LOGARITHM or to the base 10 logarithm (the effect is the same no matter which one is used). A logarithmic transformation tends to create greater dispersion among tightly clustered data, “linearizing” the scatterplot. The logarithmic transformation linearizes the relationship of economic development to democracy.

Although Figure 1 does not show the actual scatter of points, the curve captures the shape of the distribution of points better than does the straight (linear) line. Statistically expressed, the improvement in fit of the OLS regression model is notable. The R -SQUARED statistic for the linear specification is .27, while the R^2 statistic for the logarithmic specification (X logged) is .30. (The data set has $N = 130$ countries

and excludes two categories of OUTLIER countries: members of OPEC (Organization of Petroleum Exporting Countries) and countries that rely on petroleum exports for more than 50% of their export earnings.)

The other two transformations most commonly employed are the parabolic transformation and the hyperbolic transformation. A parabolic transformation is suggested if the relationship, when graphed, is one of an upside-down U-shaped curve. It involves squaring the independent variable and including it with the untransformed variable in the OLS regression model. The hyperbolic transformation is recommended if there is a threshold for the dependent variable. To take this transformation, researchers should perform the arithmetic operation $1/X$, with X representing the independent variable. These three remedies will assist the researcher in meeting the linearity assumption when the untransformed variables have a curvilinear relationship.

—Ross E. Burkhardt

REFERENCES

- Burkhardt, R. E., and Lewis-Beck, M. S. (1994). Comparative democracy: The economic development thesis. *American Political Science Review*, 88, 903–910.
- Draper, N. R. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley.
- Jackman, R. W. (1973). On the relationship of economic development to political performance. *American Journal of Political Science*, 17, 611–621.
- Kruskal, J. B. (1968). Transformation of data. In D. L. Sills (Ed.), *International encyclopedia of the social sciences* (Vol. 15, pp. 182–192). New York: Macmillan.
- Studenmund, A. H. (1992). *Using econometrics: A practical guide* (2nd ed.). New York: HarperCollins.

D

DANGER IN RESEARCH

Researchers potentially face a variety of hazards while conducting fieldwork. These might include accidents, arrest, assault, illness, harassment, and verbal abuse. These risks are probably most common for researchers working in remote locations, in situations of violent social conflict, with some kinds of deviant groups or those employed in dangerous occupations. Female researchers may face risks of sexual assault or sexual harassment. The risks posed by fieldwork shape research agendas by deterring researchers from investigating particular topics or working in particular regions. For example, the dangers involved have probably inhibited research on the violent conflicts occurring in some countries and in some regions. Fieldwork hazard is also strongly implicated in the ethics and politics of field situations, and has an impact on professional practice.

Broadly, dangers are of two kinds—ambient and situational. Ambient danger arises when a researcher is exposed to otherwise avoidable dangers from having to operate in a dangerous setting in order for the research to be carried out. For example, to study the policing of a violent social conflict, a researcher might have to accompany police officers into areas where they are vulnerable to attack. Situational danger arises when the researcher's presence or activities evoke aggression, hostility, or violence from those within the setting. Drug users, for instance, have been known to exhibit aggression toward researchers when under the influence of certain drugs. Researchers working in dangerous settings are probably most vulnerable

early on in their field experience, when they still lack situational awareness that makes setting members competent at assessing potential dangers. Research in dangerous situations often depends on the development of an interactional “safety zone,” a socially maintained area that extends for a short distance around the field worker and within which the researcher can feel secure and psychologically at ease. Establishment of a safety zone depends in part on the social acceptance of the researcher by those within the setting, as well as on the researcher's ability to size up the possible hazards involved in entering or staying in a particular setting. For example, researchers working with drug abusers might need to be alert to the possibility of fights or disputes erupting around them. They might also need to exercise caution with respect to drive-by shootings and police raids.

Personal style and demeanor can be important in some kinds of potentially dangerous situations. Researchers might find it appropriate to avoid dressing or acting in ways that single them out for undue attention and may also need to adopt a self-confident and determined demeanor to avoid denoting potential victim status to those in the setting. Failure to do so has sometimes resulted in researchers being attacked. Being socially “sponsored” by someone well-known and trusted in the setting can help to avoid danger. In potentially hazardous settings, such individuals can warn the researcher of imminent danger or draw attention to volatile or threatening situations. Some researchers, such as anthropologists who work in remote areas, may face particular hazards. There can be a risk of contracting infectious and parasitic diseases. Falls and vehicle accidents have been

identified as common causes of death and injury among anthropologists.

It should not be assumed that research in dangerous settings is impossible. Indeed, the risks involved can sometimes be exaggerated. Fieldwork dangers can be negotiated, but they need to be approached with foresight and planning. It is important to evaluate the possibility of danger, its potential sources, and how it might be managed or exacerbated by the researcher's actions. Traditionally treated rather lightly, often as a source of "war stories," the risks involved in fieldwork are now being discussed more often. Concerns about insurance liability and legal responsibilities toward employees and students have encouraged a more explicit acknowledgment of the risks involved in fieldwork. Institutional responses to fieldwork dangers now increasingly take the form of risk assessments, precautionary guidelines, policy frameworks that specify good practice, managerial responsibilities, and training provision. It is important that researchers are aware of, and contribute to, the development of such responses.

—Raymond M. Lee

See also ETHICAL PRINCIPLES, FIELD RELATIONS, PARTICIPANT OBSERVATION, POLITICS OF RESEARCH

REFERENCE

Lee, R. M. (1995). *Dangerous fieldwork*. London: Sage.

DATA

Data are the recorded empirical observations on the CASES under study, such as the annual percentage changes in the gross domestic product, the number of children in each family, scores on an attitude scale, fieldnotes from observing street gang behavior, or interview responses from old-age pensioners. Social scientists gather data from a wide variety of sources, including experiments, surveys, public records, historical documents, statistical yearbooks, and direct field observation. The recorded empirical observations can be qualitative as well as quantitative data. With quantitative data, the empirical observations are numbers that have intrinsic meaning, such as income in dollars. Qualitative data may be kept in text form, or assigned numerical codes, which may represent an efficient way to record and summarize them. For

example, in an anthropological study, a researcher may record whether the market vendor brought produce to the market by cart, animal, motor vehicle, or hand. The researcher may go on to code these observations numerically, as follows: cart = 1, animal = 2, motor vehicle = 3, hand = 4. This coding facilitates storage of the information by computer. Increasingly, computer-assisted qualitative data analysis software is being used. Finally, it should be noted that the word *data* is plural.

—Michael S. Lewis-Beck

DATA ARCHIVES

Data archives are resource centers that acquire, store, and disseminate digital DATA for secondary analysis for both research and teaching. Their prime function is to ensure long-term preservation and future usability of the data they hold.

Data archiving is a method of conserving expensive resources and ensuring that their research potential is fully exploited. Unless preserved and documented for further research, data that have often been collected at significant expense, with substantial expertise and involving respondents' contributions, may later exist only in a small number of reports that analyze only a fraction of the research potential of the data. Within a very short space of time, the data files are likely to be lost or become obsolete as technology evolves.

HISTORY

The data-archiving movement began in the 1960s within a number of key social science departments in the United States that stored original data of survey interviews. The movement spread across Europe, and in 1967, the UK Data Archive (UKDA) was established by the National Research Council for the Social Sciences. In the late 1970s, many national archives joined wider professional organizations: the Council of European Social Science Data Archives (CESSDA) and the International Federation of Data Organizations (IFDO). These were established to promote networks of data services for the social sciences and to foster cooperation on key archival strategies, procedures, and technologies.

The first data archives collected data of specific interest to QUANTITATIVE RESEARCHERS in the social

sciences. In the 1960s, these were largely opinion poll or election data, but as the trend for large-scale surveys grew, by the late 1970s, the UKDA began to acquire major government surveys and censuses. Because of their large sample sizes and the richness of the information collected, these surveys represent major research resources. Examples of major British government series include the General Household Survey, the Labor Force Survey, and the Family Expenditure Survey. In the United States, key government survey series include the Survey of Income and Program Participation, the Current Population Survey, and the National Health Interview Survey.

By the 1990s, the UKDA collection had grown to thousands of data sets spanning a wide range of data sources. Well-known U.K.-based academic repeated large-scale surveys include the British Social Attitudes Survey, the British Household Panel Survey, and the National Child Development Study. In the United States, the General Social Survey, the Panel Study of Income Dynamics, and the National Longitudinal Survey of 1972 are examples of long-running survey series. Major cross-national series include the World Values and European Values Surveys, the Eurobarometer Surveys, and the International Social Survey Program Series.

In the 1990s, the research community recognized the needs of qualitative researchers by funding a Qualitative Data Archive (Qualidata) at the University of Essex in 1994, and in 1995, a History Data Service (HDS) was established devoted to the archiving and dissemination of a broad range of historical data.

DATA HOLDINGS

Social science data archives acquire a significant range of data relating to society, both historical and contemporary, from sources including surveys, CENSUSES, registers, and aggregate statistics.

Data are created in a wide variety of formats. Data archives typically collect numeric data, which can then be analyzed with the use of statistical software. Numeric data result when textual information (such as answers to survey questions) has been coded, or they may represent individual or aggregated quantities, such as earned income.

The demand for access to digital texts, images, and audiovisual material has meant that material from in-depth interviews, FIELDNOTES to recorded

interviews, and open-ended survey questions are now available for computer analysis.

Finally, data sets from across the world are available through national data archives' reciprocal arrangements with other national data archives.

ACQUIRING DATA

A key concern for a data archive is to ensure that the materials it acquires are suitable for informed use and meet demand. All materials deposited are selected and evaluated, and they must meet certain criteria, such as being documented to a minimum standard. Furthermore, acquisitions policies must be flexible and responsive to changes in both data and information needs of the research communities, as well as in the rapidly changing climate of technology.

LONG-TERM PRESERVATION

It is the responsibility of data archives to keep up with technological advances by monitoring hardware and software developments and migrating their collection accordingly. When technology changes, the data in their holdings are technically transformed to remain readable in the new environment. Computer programs are maintained to allow data to be easily transformed from an in-house standard to the various formats required by users.

PREPARING DATA FOR ARCHIVING

On acquisition, data archivists undertake data-processing activities. First, the data set is checked and validated, for example, by examining numeric data values and ensuring that data are anonymous so that the risk of identifying individuals is minimal. Second, "metadata" (data about data) are produced with the aim of producing high-quality finding aids and providing good user documentation. Metadata cover information describing the study and the data and include enhancements added to a numeric data file (such as creating variable and value labels that enable users to view variable information and frequencies). A systematic catalogue record is always created for studies, detailing an overview of the study, the size and content of the data set, its availability, and terms and conditions of access. User guides further contain information on how the data were collected, the original QUESTIONNAIRES, and how to use the data.

PROVIDING ACCESS TO DATA

Data supplied by data archives can be used for reporting of statistics, such as basic frequency counts, and for more in-depth secondary analysis. Secondary analysis strengthens scientific inquiry, avoids duplication of data collection, and opens up methods of data collection and measurement. Reusing archived data enables new users to ask new questions of old data; undertake COMPARATIVE RESEARCH, REPLICATION, or restudy; inform RESEARCH DESIGN; and promote methodological advancement. Finally, data can provide significant resources for training in research and substantive learning.

Users typically request data in a particular format, such as a statistical or word-processing package. These days, data can be accessed via instant Web download facilities or can be dispatched on portable media such as CD-ROM. The 21st century has seen a move toward sophisticated online analysis tools, where users can search and analyze subset data via a Web browser, such as NESSTAR.

Data supplied by data archives are normally anonymous to ensure that the risk of identifying individuals is minimal. Users are typically required to sign an agreement to the effect that they will not attempt to identify individuals when carrying out analyses.

—Louise Corti

REFERENCES

- ICPSR. (2002). *Guide to social science data preparation and archiving* [online]. Available: http://www.ifdo.org/archiving_distribution/datprep_archiving_bfr.htm.
- IFDO. (n.d.). [online]. Available: <http://www.ifdo.org/>.
- Ryssevik, J., & Musgrave, S. (2001). The social science dream machine: Resource discovery, analysis, and delivery on the Web. *Social Science Computing Review*, 19(2), 163–174.
- UK Data Archive (n.d.). [Online]. Available: <http://www.data-archive.ac.uk/>.

DATA MANAGEMENT

We need to manage research data in a way that preserves its utility over time and across users. This means we must pay attention to data documentation and storage so that, at a minimum, others can access our data; replicate our analyses; and, preferably, extend them. How we achieve these goals depends upon

whether we produce the data (primary data collection) or simply reuse it (secondary data analysis; see SECONDARY DATA, SECONDARY ANALYSIS OF SURVEY DATA), as well as what kind of data we are dealing with.

As we discuss managing social science data, we distinguish between the two most common types of data: SURVEY data and records data. Social scientists deal with other types, such as data from EXPERIMENTS and audiovisual recordings. Audio or video recordings create their own data management problems, and people in the broadcast industry likely have better insights to management and archival issues, so we will not attempt to give advice in this area. Increasingly, economists collect experimental data from laboratory settings to learn how economic agents react to controlled economic opportunities. These experiments involve customized software that will, itself, generate the data for analysis. In general, many of the same principles that apply to survey data apply to experimental data, but the applications are so specialized that we will say little more about experimental data.

The primary data collector bears the most onerous and important data management responsibilities; in order for others to exploit the data, the collector must successfully execute the primary data management task. The data management task begins with a plan that encompasses the entire project from data collection through analysis. Naturally, the sort of data collected dictates the plan for data management. For secondary data, the management task begins downstream from collection, and the scope of the data management problem crucially depends on the sorts of data resources that the collector has passed along to the user. It is likely that the best strategy for the data manager of secondary data is to start with the data system used by the primary data collector.

For a variety of data management applications, the relational database technology dominates other strategies. Most social scientists are not familiar with how these systems work and therefore tend to manage their data within software systems with which they are familiar, such as SAS, SPSS, or Stata. For small projects, this may be the only feasible approach; it does not make sense to learn a new technology that might be used for only a few months, unless one foresees a career requiring strong data management skills.

We will begin our discussion of data management strategies with records data. These data are especially well suited to relational databases, and applications that exemplify the technique are easier to understand

in the context of records data. Although survey data are perhaps more widely used, the reader will probably better understand how relational database methods can be used with survey data after working through an example of research data derived from administrative records.

ADMINISTRATIVE RECORDS DATA

Records data come from many sources and almost always originate with someone other than the researcher. The appropriate data management structure with records data depends upon the complexity of the data. Typical records data have a simple structure and can be maintained within the statistical package used for analysis. For example, national income account and other macroeconomic data have a fixed set of variables that repeats for various time periods. Gross national product data are quarterly, unemployment rates are monthly, and so forth. Data on stock transactions are available in a variety of temporal aggregations and even disaggregated as the full series of transactions.

However, some records data can be very complex. For example, states have massive databases on earnings reported for unemployment insurance purposes, tax returns, and people receiving benefits from transfer programs such as TANF (the former cash assistance for poor families with children, a program that has a variety of state-specific names following the reform of the welfare system in the late 1990s), Medicaid, and food stamps. These data cover both individuals and families with records from applications, payments, and administrative actions.

Let us consider the example of transfer program administrative data to describe how a relational database strategy allows the researcher to organize the data for analysis. We will simplify the example for exposition, so the reader should not infer that any particular state organizes its data in the manner we describe.

Relational databases consist of a set of tables, each of which contains a fixed set of variables. A relational database does not require the tables to be related in any particular way, nor does it require the data to be combined and extracted in any particular way. If a table contains data on a person (or company, transaction, etc.), we will use the term *row* to refer to the set of variables for that table that applies to that particular person (or company, transaction, etc.). This flexibility makes relational databases excellent

candidates for archiving and retrieving complex data from a wide range of applications. Because household size and income determine the size of welfare payments, one likely table in a database for welfare data would describe the person receiving payments under a particular program. Let us use the term *assistance group* to describe the set of people eligible for benefits under a particular program. In a single household, a person may be part of one assistance group for TANF purposes and another group for food stamp purposes. Our hypothetical state will, as a result, have data describing each assistance group for a type of benefit, with each assistance group having a unique identifier. A TANF assistance group data set might have a table that identifies the name; address; demographic characteristics (date of birth, ethnicity, sex, citizenship, etc.); Social Security number; and possibly a state identification number for the person receiving benefits on behalf of each assistance group. Another set of tables would document the checks issued to the assistance groups, with one record containing the assistance group identifier, the date a payment was made, the size of the payment, and indicators for any special administrative details relating to the payment. Similarly, there might be a table that stores the outcome of eligibility reviews or records from visits by a caseworker.

Another table in the database might relate to individuals. For example, for each month that a person generates a payment to an assistance group, there might be a row in the table that contains the person's name, date of birth, sex, ethnicity, disability status, the program under which a payment was made, and the identifier of the assistance group to which the payment was made for the benefit.

States also maintain the database showing quarterly earnings covered by the unemployment insurance program. A table in this database would contain the Social Security number of the worker, the Employer Identification number for the firm, the quarter for earnings, and the amount of the worker's earnings.

From this latter set of records, we could construct a quarterly history of the earnings for any person in the state. Similarly, we could construct a month-by-month history for any person in the state that showed in which months the person generated food stamp payments or TANF payments. It is also possible to create an event history of any person's membership in an assistance group and how that relates to earned income. We could also use these data to determine whether the child becomes the beneficiary of benefits as a part of a new

assistance group when a child's parent loses welfare benefits. Data from different tables can be *joined* together by matching rows that share a common value of some identifier, such as a Social Security number. *The advantage of a relational database is not that it is in the right format for any particular analysis, but that it allows the researcher to combine data elements together into the right format for virtually any analysis.* This flexibility explains its dominance for large commercial applications that range from e-commerce Web sites to accounting and inventory systems. Frequently, companies integrate *all* their record systems within a collection of tables of a large relational database, providing extraordinary power to track activity within the enterprise. In a well-designed relational database, all redundant data items are eliminated. This reduces both maintenance costs and retrieval errors.

When researchers know exactly what data they need and in what format it must be presented, they are likely to extract from records data only the variables they think they need and to move the data into a form convenient for their analysis strategy (e.g., readable by their statistical package). Such researchers are unlikely to use a relational database, and if the project develops as expected, this will be the right decision. However, whatever work such researchers have done with the primary data may be of little value to others using the data who have a different vision of which data are needed and how they should be organized. This violates the basic goal we set out for data management: Preserve the utility of the data across time and across users. Relational databases may not be the ideal data management system for any particular application, but they will be very good systems for almost any application.

SURVEY DATA

Survey data often possess a complex structure. Moreover, survey data are often collected with secondary analysis in mind, and for the most influential surveys, secondary analysis is the primary objective. Surveys are natural applications for relational database techniques because they are frequently made up of sets of tables of variables that are related to one another in a variety of ways.

The samples themselves most often come from multistage STRATIFIED random SAMPLES rather than the SIMPLE RANDOM SAMPLING we see in textbooks. In addition, we often interview multiple people from a unit chosen from the SAMPLING FRAME. It is essential

that the data file preserve the relationships between the data and the sampling structure and respondent relationships. For example, in area PROBABILITY SAMPLING, this means retaining for each observation the identification of the primary sampling unit and substratum, including census tract and block group identifiers, along with the probabilities of selection at each step in the sampling process. These geographic attributes define the way in which various observational units (respondents) are connected to one another. The recent explosion in the use of Geographic Information Systems (GIS) data argues strongly that the data collector capture the longitude and latitude for dwelling units either using a global positioning satellite receiver or by geocoding the street address to latitude and longitude using a package such as Maptitude or ArcView. Latitude and longitude data are the basis for organizing most GIS data, so these should be the primitives of location, not block group, census tract, or some other indexing system.

Survey data frequently utilize complex questionnaires, and this creates serious data management problems. Indeed, as computer-assisted techniques dominate the industry, instruments that heretofore would have been impossible for interviewers to administer have become the rule. We encourage the reader interested in survey data to also read the entry on COMPUTER-ASSISTED PERSONAL INTERVIEWING, which deals with many of the same issues and, together with this entry, provides a holistic view of the survey data management process from design through data dissemination.

Instruments often collect lists, or rosters, of people, employers, insurance plans, medical providers, and so on and then cycle through these lists asking a set of questions about each person, employer, insurance plan, or medical provider in the roster. These sets of related answers to survey questions constitute some of the tables of a larger relational database where the connections among the tables are defined by the design of the questionnaire.

One can think of each question in a survey as a row within a table with a variety of attributes that are linked in a flexible manner with other tables. The attributes (or "columns") within a question table would be the following:

- The question identifier
- Descriptors that characterize the content of the question (alcohol use, income, etc.)

- The question text
- A set of questions or check items that leads into the question
- A set of allowable responses to the question and data specifications for these allowable responses (whether the answer is a date, time, integer, dollar value, textual response, or a numerical value assigned to a categorical response, such as 1 = yes, 0 = no)
- Routing instructions to the next question or check item that are contingent on the response to the current question
- Instructions to assist the interviewer and respondent in completing the question
- Alternate language versions for question attributes that are in text form
- Comments about the accuracy or interpretation of the item or its source

We often refer to these attributes of questions as “metadata.” In complex surveys, these pieces of information that describe a question are connected to that question. For example, metadata include which questions lead into a particular question, and questions to which that question branches. These linkages define the flow of control or skip pattern in a questionnaire. With a sophisticated set of table definitions that describes a questionnaire, one can “join” tables and rapidly create reports that are CODEBOOKS, questionnaires, and other traditional pieces of survey documentation.

If we break down the survey into a sequence of discrete transactions (questions, check items, looping instructions, data storage commands, etc.) and construct a relational database, with each transaction being a row in a database table and the table having a set of attributes as defined in the relational database, we can efficiently manage both survey content and survey data.

Relational database software is a major software industry segment, with vendors such as Oracle, Sybase, IBM, and Microsoft offering competitive products. Many commercial applications use relational database systems (inventory control; accounting systems; Web-based retailing; administrative records systems in hospitals, welfare agencies, and so forth, to mention a few), so social scientists can piggyback on a mature software market. Seen in the context of relational databases, some of the suggested standards for codebooks and for documenting survey data, such as the data documentation initiative, are similar to relational database designs but fail to use these existing professional tool sets and their standard programming conventions.

The primary data collector has several data management choices: (a) Design the entire data collection strategy around a relational database, (b) input the post-field data and instrument information into a relational database, or (c) do something else. For simple surveys, the third alternative likely translates into documenting the survey data using a package like SAS, SPSS, or Stata. These are excellent statistical packages, and one can move data among them with a package like Stata’s Stat/Transfer. Statistical packages are, themselves, starting to incorporate relational database features. For example, SAS supports SQL (standard query language) queries to relational databases, and it also connects to relational databases. The trend for many years has been toward relational databases to manage databases, starting with the largest, most complex projects. When setting up large projects, social scientists would do well to build their data management strategies and staff around relational database management systems.

—Randall J. Olsen

See also QUALITATIVE DATA MANAGEMENT

REFERENCES

- Elmasri, R. A., & Navathe, S. B. (2001). *Fundamentals of database systems*. New York: Addison-Wesley.
- Gray, J., & Reuter, A. (1992). *Transaction processing: Concepts and techniques*. San Francisco: Morgan Kaufmann.
- Kroenke, D. M. (2001). *Database processing: Fundamentals, design and implementation*. Upper Saddle River, NJ: Prentice Hall.
- Stern, J., Stackowiack, R., & Greenwald, R. (2001). *Oracle essentials: Oracle9i, Oracle8i and Oracle 8*. Sebastopol, CA: O’Reilly.

DEBRIEFING

Originally a military term, *debriefing* means questioning or instructing at the end of a mission or period of service. As used in human research, debriefing refers to a conversation between investigator and subject that occurs after the research session. Debriefing is the post-session counterpart of INFORMED CONSENT and should be conducted in a way that benefits and respects the subject.

Debriefing may have several purposes. Generally, it is an opportunity for the subject to ask questions and for the investigator to thank the subject for

participating, more fully explain the research, and discuss the subjects' perception of the research experience. Research participation can have educational or therapeutic value for participants, and debriefing is an appropriate opportunity to consolidate these values through appropriate conversation and handouts. Debriefing also provides an important opportunity for the researcher to learn how subjects perceived the research and why they behaved as they did. The perceptive researcher may learn as much from the debriefing as from the research itself. For these values of debriefing to be realized, it is essential that the researcher prepare appropriate information to be conveyed, provide a respectful opportunity for the subject to ask questions and express reactions, and allow an unhurried period in which the interested subject and appreciative investigator can interact. The debriefing should be planned and scheduled as an integral part of the research, and it should be appropriate to the circumstances. For example, if the research was about a private or sensitive matter, the debriefing should occur privately. In the case of research on children or on adults of limited autonomy, the subject's parent or guardian, as well as the subject, should be debriefed appropriately.

Sometimes, investigators promise to send subjects the results of the study rather than discussing the study with them after their session. There are many problems with this arrangement. Such promises are often broken. Completion of the research typically occurs so much later that it may be impossible to re-contact the subjects; if re-contact occurs, the information arrives out of context and after the experience has been forgotten. Besides, the results of a single study are often meager and of limited interest to subjects. Of greater value is the knowledge that the researcher gained from the literature search that (should have) preceded the study and that can easily be summarized in layman's language during the debriefing.

When the research involves concealment or DECEPTION (any procedure that causes the subject to believe something that is not so), the debriefing should include explaining the nature of the deception and why it was employed (*dehoaxing*). After a deceptive or stressful session, the debriefing should also include *desensitizing*, an effort to return the subject to a positive emotional state. Deception is easy to debrief when honest, straightforward induction involves having subjects

- Consent to participate in one of various specified conditions (as in placebo research)

- Consent to deception
- Waive their right to be informed (see Sieber, 1992, for a discussion of these methods)

Such subjects expect to have the real situation revealed in a debriefing. In contrast, subjects who were falsely informed may react negatively to the revelation that they have been fooled, and they may wonder whether they can believe the dehoaxing. Desensitizing, under such conditions of mistrust, may be impossible. Similarly, subjects who do not know they have participated in research may cling to the belief that they had engaged in a naturally occurring (uncontrived) situation, and may be difficult to debrief. Whatever device was used to deceive (e.g., a confederate, misrepresentation of an object or procedure), a convincing demonstration of the deception is usually easy to arrange. For example, if false feedback was given on a test performance, the dehoaxing could include giving subjects their own tests, still in sealed envelopes as subjects had submitted them.

WHEN NOT TO DEHOAX

Dehoaxing may involve explanation of the research design and procedures, including the hypotheses and the comparisons that were being made. However, debriefing should be appropriate to the culture and level of sophistication of the subjects and should not involve blaming, demeaning, or confusing subjects. Hence, sophisticated details may be inappropriate, as in situations such as the following in which subjects most likely would be unable to understand or benefit from a detailed dehoaxing:

- A study of children's ability to resist temptation to take (steal) items that do not belong to them, as a function of stage of moral development
- A study of child-rearing practices and child aggression that compares populations of parents who espouse corporal punishment with parents who espouse other approaches
- A study of explanations of illness by members of cultures that tend to invoke explanations involving evil spirits or witchcraft

It might be appropriate to provide a simple, non-judgmental description of the research, such as, "We wanted to learn the various methods parents use to discipline their children. We appreciate your letting us observe how you interact with your child." If

subjects express a desire to know more, additional nonjudgmental information could be offered, such as, "Here are some of the kinds of methods we observed."

Designing appropriate debriefing requires sensitivity to the needs, culture, and feelings of subjects. The main purposes of debriefing are to thank subjects for participating; answer their questions; learn their reactions; and, in a respectful way, convey honest and perhaps educationally beneficial information to subjects.

—Joan E. Sieber

REFERENCES

- American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist*, 47, 1597–1611.
- Holmes, D. (1976). Debriefing after psychological experiments: Effectiveness of post-experimental desensitizing. *American Psychologist*, 32, 868–875.
- Sieber, J. E. (1992). *Planning ethically responsible research*. Newbury Park, CA: Sage.

DECEPTION

Deception takes the form of a lapse or calculated misrepresentation in the process of informing participants that research is taking place and of its nature, purpose, and consequences.

Some measure of deception is widely practiced in social research, although it is not always identified as such. The story given to participants to inform their consent is often partial; investigators fear that they will lose willing respondents if they reveal the whole truth. They also use misleading descriptions of their interests, for example, talking of recreation and health when they mean sex and drugs.

In the 1960s and 1970s, however, before social researchers had developed a collective conscience, deception was rather more blatant; it often took the form of the adoption of a role, disguise, or false identity by investigators entering a research field. For example, Rosenhan (see Bulmer, 1982) and his collaborators simulated the symptoms of insanity in order to gain admission to mental institutions and test screening procedures. Howard Griffin took a medication to change the color of his skin in order to conduct a pseudo-participant study of being black in the American Deep South; contrary to expectation, this procedure proved to be irreversible. In order to secure the cooperation of

a pastor of an Afro-Caribbean congregation in Bristol, Ken Pryce signified beliefs he did not hold and allowed himself to be baptized (Homan, 1992).

Stanley Milgram's infamous studies of obedience and authority involved deception of a different kind. Milgram misled those he hired into thinking they were his assistants; they were then inveigled into believing that they were inflicting pain on other participants, who were, in reality, trained actors. Subjects suffered various adverse aftereffects as they came to terms with their willingness to cause pain and distress.

Milgram's case and others have prompted concerns about the effect deception has upon those who practice it and upon the reputation of research professionals. At the individual level, deceit and betrayal may become habitual, and distrust endemic. For the community of researchers, there is a risk that they gain a reputation for deviousness and fraudulence. The maintenance of an unfamiliar role puts considerable strain upon those who are trained in social science rather than in drama school. A change of color, residence in a mental hospital, or the company of a criminal gang are not undertakings one would wish upon oneself, let alone prescribe for one's students.

But whether deception in social research is always morally wrong is an open question. We sympathize with participants in the obedience experiments who suffered nervous aftereffects. We turn on those who conducted these projects and reckon that a general taboo on deception would have forestalled these consequences. But when deception is used to bypass the barriers of powerful or socially disapproved groups like the Ku Klux Klan, the outrage is rather muted. In the instance of groups that can look after themselves and resist conventional forms of inquiry, there may be an argument for deceptive strategies. Researchers must consider the conflicting responsibilities they have to their subjects and to the publication of truth.

—Roger Homan

REFERENCES

- Bulmer, M. (Ed.). (1982). *Social research ethics*. New York: Holmes and Meier.
- Homan, R. (1992). *The ethics of social research*. London: Longman.
- Lauder, M. (2003). Covert participant observation of a deviant community: Justifying the use of deception. *Journal of Contemporary Religion*, 18(2), 185–196.
- Milgram, S. (1974). *Obedience to authority*. London: Tavistock.

DECILE RANGE

The decile range is calculated in a similar way, and has a similar role, to the INTERQUARTILE RANGE. A frequency distribution is arrayed from low to high and divided into 10 equal groups in terms of the number of cases in each group. The first and last decile are removed, and the decile range is the difference between the lowest and the highest values that are left. As with the interquartile range, the decile range is a way of dealing with the problem of OUTLIERS that affects the RANGE. The term is sometimes used, though incorrectly and inaccurately, as a synonym of “decile.”

—Alan Bryman

DECONSTRUCTIONISM

Deconstructionism is concerned with identifying and problematizing the categories and strategies that produce particular versions of reality. Texts are taken apart to observe how they are constructed in such a way as to exhibit particular images of people, places, and objects. French social theorist Jacques Derrida (1976) first introduced deconstructionism to philosophy in the late 1960s as part of the poststructuralist movement. Although closely linked to literary theory, deconstructionism has also found resonance in the study of architecture, art, culture, law, media, politics, and psychology.

Rooted in close readings of a text, deconstructionism enables us to identify how the construction of the text is dependent on contradictions, confusions, and unstated absences. Derrida objects to the logic of fixed categories that have dominated Western thought and philosophy, especially the either/or logic of binary oppositions such as mind/body, masculine/feminine, health/illness, and individual/society. Instead of the either/or logic of binary oppositions, Derrida argues that we should accept the logic of both/and. Phenomena are best understood when we consider what they appear to both include and exclude. For example, the individual/society opposition can be alternatively understood as conjoined features of a system, where each component requires the other to exist and to make sense (Burr, 1995). Oppositions serve as devices to highlight ethical, ideological, and political processes at work in a text.

Ian Parker (1988) recommends three steps or stages to deconstructionist analysis. Step 1 involves identifying BINARY oppositions or polarities in the text and illustrating the way in which one particular pole is privileged. Step 2 shifts the subordinate pole of the opposition into the dominant position. This serves to demonstrate how the privileged pole depends on and could not exist, and would not function, without the other opposition. Finally, in Step 3, the opposition is reconsidered, and new concepts and practices are developed. Parker has used these stages of deconstructionism when analyzing the discourse of a British radio soap opera, *The Archers*. Through an analysis of transcripts of this soap opera, Parker contests the distinctions between reality and fiction, and everyday explanations and theoretical explanations that are used when discussing gender relations.

Deconstructionism is controversial. Derrida's work has been praised for making one of the most significant contributions to contemporary thinking while simultaneously being denounced as corrupting widely accepted ideas about texts, meanings, and identities. Proponents of deconstructionism argue that it provides a novel and stimulating approach to the study of language and meaning, and that it can disrupt theories and serve to open up debate and conflicts. Equally, deconstructionism can serve to change power relations and can encourage subversion of dominant ideas. Opponents, however, claim that deconstructionism is concerned with verifying that which cannot be verified or proven. Verification through deconstructionism is immensely difficult given that its own object of inquiry is unstable. It is impossible to locate the meaning of a text because it consists of an endless chain of signification. Therefore, deconstructionist analysis is limited to merely exemplifying meaning in a text. Deconstructionism has also been criticized for prioritizing the role of texts when understanding meaning making and power struggles, to the detriment of considering the role of the audience.

—Sharon Lockyer

REFERENCES

- Burr, V. (1995). *An introduction to social constructionism*. London: Routledge.
- Derrida, J. (1976). *Of grammatology* (G. Spivak, Trans.). Baltimore, MD: Johns Hopkins University Press.
- Parker, I. (1988). Deconstructing accounts. In C. Antaki (Ed.), *Analysing everyday explanation: A casebook of methods* (pp. 184–198). London: Sage.

DEDUCTION

In logic, deduction is a process used to derive particular statements from general statements. The logic is used in the social sciences to test THEORY. A HYPOTHESIS is deduced from a theory and is tested by comparison with relevant DATA. If the data do not agree with the hypothesis, the theory is rejected as false. If the data agree with the hypothesis, the theory is accepted provisionally; it is corroborated. Testing such a hypothesis is assumed to be a test of the theory itself.

Notions of deduction go back to antiquity, to Euclidean geometry and Aristotelian logic. Deduction is an essential aspect of the HYPOTHETICO-DEDUCTIVE METHOD, which is associated with Karl Popper's FALSIFICATIONISM. Popper advocated the logic of deduction to overcome the deficiencies of INDUCTION. The core of his argument is that because observations, or data, do not provide a reliable foundation for scientific theories, and because inductive logic is flawed, a different logic is needed for developing theories. His solution was to accept that all data collection is selective and involves interpretation by the observer, and then to use a logic that is the reverse of induction.

Instead of looking for confirming evidence to support an emerging generalization, as occurs in induction, Popper argued that the aim of science is to try to refute the tentative theories that have been proposed. The search for truth is elusive because we have no way of knowing when we have arrived at it. All that can be done is to eliminate false theories by showing that data do not fit with them. Theories that survive the testing process are not proven to be true because it is possible that further testing will find disconfirming evidence. It may be necessary to discard a theory, or at least to modify it and then conduct further research on it. Therefore, science is a process of conjecture and refutation, of trial and error, of putting up a tentative theory and then endeavoring to show that the theory is false.

An example of what a deductive theory might look like comes from a reconstruction of Durkheim's theory of egoistic suicide. It consists of five propositions using three main concepts: suicide rate (the number of suicides per thousand of a population); individualism (the tendency of people to think for themselves and to act independently, rather than to conform to the beliefs and norms of some group); and Protestantism (a collection of Christian religious groups formed following

the Reformation and the subsequent fragmentation of Roman Catholicism). The five propositions are as follows:

1. In any social grouping, the suicide rate varies directly with the degree of individualism (egoism).
2. The degree of individualism varies directly with the incidence of Protestantism.
3. Therefore, the suicide rate varies with the incidence of Protestantism.
4. The incidence of Protestantism in Spain is low.
5. Therefore, the suicide rate in Spain is low (Homans, 1964, p. 951).

The conclusion to the argument becomes the hypothesis to be tested. If the hypothesis is refuted, the theory is assumed to be false. By changing the country, for example, predictions can be made and the hypothesis tested in other contexts.

Like inductive logic, deductive logic has been severely criticized. Some major criticisms are the following: If observations are interpretations, and we can never observe reality directly, regularities cannot be established confidently nor theories refuted conclusively; the tentative acceptance of a yet-unrefuted theory requires some inductive support; paying too much attention to logic can stifle scientific creativity; and the process of accepting or rejecting theories involves social and psychological processes, not just logical ones (see Blaikie, 1993, and Chalmers, 1982, for critical reviews).

—Norman Blaikie

REFERENCES

- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Chalmers, A. F. (1982). *What is this thing called science?* St. Lucia: University of Queensland Press.
- Homans, G. C. (1964). Contemporary theory in sociology. In R. E. L. Faris (Ed.), *Handbook of modern sociology* (pp. 951–977). Chicago: Rand McNally.
- Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

DEGREES OF FREEDOM

The number of independent observations available for PARAMETER ESTIMATION indicates the degrees

of freedom, df . Generally speaking, a test statistic has degrees of freedom, determined by mathematical constraints on the quantities to be estimated. As examples, degrees of freedom must be calculated with the T TEST, the F RATIO, the CHI-SQUARE, the STANDARD DEVIATION, or REGRESSION analysis. In the case of regression analysis, that number may be calculated by subtracting from the sample size the number of coefficients to be estimated, including the constant.

Consider the simple case of estimating the STANDARD DEVIATION of a population variable, from a random sample of N observations. The formula is as follows:

$$SD = \sqrt{\frac{\sum (U - \bar{U})^2}{N - 1}},$$

where SD = the standard deviation estimated in the sample, $\sqrt{}$ = square root, $\sum$ = sum, U = sample observations on the population variable, $\bar{U}$ = the average score of the sample observations on U , and N = the sample size. Note that the numerator is divided by $N - 1$, not N . Substantively, this adjustment may appear small. Nevertheless, it is theoretically important. The correction is necessary to avoid an exaggerated, BIASED estimate of the population standard deviation. In calculating this 1 population parameter, we have exhausted 1 degree of freedom. This is so because of the characteristics of the sample mean, used in the formula. There are N number of $(U - \bar{U})$ scores. However, just $N - 1$ of those are independent because once they are computed, the last, the N th value, is necessarily fixed. This comes from the fact that always $\sum (U - \bar{U}) = 0$. For instance, suppose $N = 3$, and $(U_1 - \bar{U}) = 6$, $(U_2 - \bar{U}) = 12$. Then, it must be that $(U_3 - \bar{U}) = -18$. The restriction on the deviations from the mean, that they sum to zero, dictates that the last deviation is not “free” but fixed, so 1 degree of freedom is lost. In the example, only one population parameter, the standard deviation, is estimated, leaving degrees of freedom = $N - 1$.

In the more complicated case of regression analysis, the $df = N - K$, where N = the sample size and K = the number of parameters to be estimated. Suppose the MULTIPLE REGRESSION model of $Y = a + bX + cZ + e$, where $N = 45$. There are three parameters to be estimated—the intercept a and the two slopes, b and c . Hence, $df = N - K = 45 - 3 = 42$. With regression analysis, it is important to have a sufficient number of degrees of freedom. Take the extreme, when the SIMPLE CORRELATION (REGRESSION)

model $Y = a + bX + e$ is fitted to a sample with $N = 2$. Here the R -SQUARED = 1.0, because the straight line of the ORDINARY LEAST SQUARES fit connects the two data points without error. This is not a perfect explanation but mere mathematical necessity, for the degrees of freedom have been exhausted, that is, $N - K = 2 - 2 = 0$. Generally, in regression analysis, one wants many more independent observations than independent variables. When K begins to approach N , analysts pay special attention to the ADJUSTED R -SQUARED, which corrects for lost degrees of freedom.

—Michael S. Lewis-Beck

REFERENCES

- Kennedy, P. (1998). *A guide to econometrics* (4th ed.). Cambridge: MIT Press.
 Lewis-Beck, M. S. (1995). *Data analysis: An introduction*. Thousand Oaks, CA: Sage.

DELETION

Deletion is a method for dealing with MISSING DATA, which are usually eliminated (or transformed) before analysis begins. The most common type of missing data deletion by computer is called LISTWISE deletion (also called casewise deletion). With listwise, a case is eliminated if it has a missing value on any of the variables under analysis. An alternative to listwise deletion is PAIRWISE deletion, where a case is eliminated only if it has missing value(s) on a pair of variables being studied. Analysis might then proceed, for example, from the estimated CORRELATION matrix. The pairwise strategy has the advantage of keeping up SAMPLE size, but has the disadvantage that it can produce statistical inconsistencies in the results.

—Michael S. Lewis-Beck

DELPHI TECHNIQUE

This is a qualitative method for obtaining consensus among a group of experts. Named after the Delphic oracle, skillful at forecasting the future in ancient times, it was invented by the RAND Corporation in the 1960s for forecasting the probability, desirability, and impact

of future events. Its use has now been extended to a wide range of situations where convergence of opinions is desirable. There are now three distinct forms: *exploratory Delphi*, as developed by RAND; *focus Delphi*, seeking views of disparate groups likely to be affected by some policy; and *normative Delphi*, gathering experts' opinions on defined issues to achieve consensus (e.g., to set goals and objectives). The essential element in the Delphi process is anonymity of participants when giving their opinion, which alleviates problems that could be caused by domination of the group by a few powerful individuals. Quantitative estimates can also be obtained (e.g., spending priorities and relative allocations of money).

The Delphi process proceeds in an iterative way as follows:

Round 1. Participants are invited to comment on a subject or a set of opinions provided by the facilitating group. Opinions should be based on participants' personal knowledge and experience. The facilitators then analyze the results and provide feedback, such as graphical summaries, rankings, or lists under a limited number of headings. The results are then fed back to participants as input into a second-round questionnaire.

Round 2. In view of the feedback based on Round 1, participants are asked to comment on the summarized views of others and also reconsider their opinions. As before, the facilitators analyze the opinions received, and these are again fed back to participants as input into the questionnaire for the next round. This process of repeated rounds continues until "adequate" convergence is obtained. Usually, three rounds are sufficient. Often, in addition to scoring their agreement with results, participants are asked to rate the degree of confidence they have in their opinions—this can be used in weighting responses to produce summary results, provided the more expert participants do not turn out to be the most modest. With the advent of the Internet, and, in the future, The Grid, there is now the opportunity to automate much of the facilitators' analyses and speed up the process.

It is important that participants be, in some sense, "experts" for valid results to be reached, and there should be sufficient participants that the extreme views of a few do not influence the results for the group disproportionately. Users of a service can be considered "experts" in what they want from a service, but professionals may be better informed as to potential risks and benefits. Both can be accommodated in the same study, with each group being asked to take into account the

other's views (e.g., see Charlton, Patrick, Matthews, & West, 1981). Consideration also needs to be given at the start as to what is fed back to participants after each round, and how "agreement" is defined, including the treatment of outliers. The validity and acceptability of the results need to be assessed at the end of the study. The Delphi method may be most appropriate when opinions are being sought, rather than in situations where evidence can be synthesized using conventional statistical methods. Delphi may be seen more as a method for structuring group communication than providing definitive answers.

—John R. H. Charlton

REFERENCES

- Charlton, J., Patrick, D. L., Matthews, G., & West, P. A. (1981). Spending priorities in Kent: A Delphi study. *Journal of Epidemiology and Community Health*, 35, 288–292.
- Linstone, H. A. (1978). The Delphi technique. In R. B. Fowles (Ed.), *Handbook of futures research*. Westport, CT: Greenwood.
- Jones, J. M. G., & Hunter, D. (1995). Consensus methods for medical and health services research. *British Medical Journal*, 311, 376–380.
- Murphy, M. K., Black, N. A., Lamping, D. L., McKee, C. M., Sanderson, C. F. B., Ashkam, J., & Marteau, T. (1998). Consensus development methods, and their use in clinical guideline development [entire issue]. *Health Technology Assessment*, 2(3).
- Pope, C., & Mayes, N. (1995). Qualitative research: Reaching the parts other methods cannot reach: An introduction to qualitative methods in health and health services research. *British Medical Journal*, 311, 42–45.

DEMAND CHARACTERISTICS

In some research, typically laboratory-based experimental research, the expectations of the researcher are transparent, the hypotheses under investigation are discernable, and the hoped-for response is obvious. Such research contexts are said to be characterized by high degrees of experimental demand characteristics. Demand characteristics are cues provided participants (by the experimenter, the research context, etc.), often unintentionally and inadvertently, regarding the appropriate or desirable behavior. A smile by the researcher when the participant has made the sought-for response, a slight nod of the head, a minor change in vocal inflection—all or any of these actions might bias the

outcome of an investigation. Demand characteristics seriously undermine the INTERNAL and EXTERNAL VALIDITY of research because they can cause outcomes that are misattributed to the experimental treatment. When variations in results are the result of demand effects, the interpretation that implicates the experimental treatment as the cause of observed results is in error.

Robert Rosenthal (1976) was most persistent in calling social scientists' attention to the distorting effects of experimental demand, and in a series of studies, he showed how experimenters' expectations could be transmitted to research participants. In one of his most controversial studies, he suggested that teachers' expectations regarding the future achievement of their pupils affected students' actual achievement gains (Rosenthal & Jacobson, 1968), and later meta-analytic research has supported this interpretation. As a result of heightened attention to the potentially biasing effects of demand characteristics, social researchers today are careful not to make obvious the hypotheses under investigation, often resorting to elaborate COVER STORIES to hide the true purpose of their research. In addition, to shield against inadvertent cueing of the sought-for responses, studies often are conducted in such a manner that neither participant nor researcher knows the particular treatment condition to which any given individual has been (randomly) assigned. Such DOUBLE-BLIND techniques are now common in social research and represent a powerful defense against results that otherwise might be tainted by demand characteristics.

Demand characteristics are related to the particular mental set adopted by the participants, as well as their self-consciousness as subjects of research. Crano and Brewer (2002) grouped participants into three general categories: voluntary participants, who are aware of, and satisfied with, their role as subjects of study; involuntary participants, who feel they have been coerced into serving in a study; and nonvoluntary participants, who are unaware that they are under investigation. Of these, voluntary participants are most likely to be affected by demand characteristics because they have made the conscious decision to participate; do not feel unduly coerced; and, thus, may be prone to try to "help" the researcher by acting so as to confirm the hypotheses that are apparently under study. Involuntary participants are more likely to attempt to sabotage the research, in that they feel they were coerced into serving as participants. Nonvoluntary participants are

not likely to attempt to help or hurt the research because they are, by definition, unaware that they are under investigation.

—William D. Crano

REFERENCE

- Crano, W. D., & Brewer, M. B. (2002). *Principles and methods of social research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Rosenthal, R. (1976). *Experimenter effects in behavioral research*. New York: Irvington.
- Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart and Winston.

DEMOCRATIC RESEARCH AND EVALUATION

Deliberative democratic research and evaluation aspires to arrive at unbiased conclusions by considering all relevant interests, values, and perspectives; by engaging in extended dialogue with major stakeholders; and by promoting extensive deliberation about the study's conclusions.

—Ernest R. House

REFERENCE

- House, E. R., & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage.

DEMOGRAPHIC METHODS

Demographic methods include a wide range of measures and techniques used to describe POPULATIONS. Although usually applied to human populations, they are increasingly used to describe nonhuman populations, ranging from animals and insects to manufactured products.

The most basic demographic techniques, such as the calculation of proportions and ratios, describe the composition of a population, such as its age-sex structure, racial composition, or geographic distribution. Measures such as rates and probabilities describe population dynamics. Some demographic methods focus on the separate processes producing change in a population. Single and multiple decrement LIFE TABLES, for example, focus on exit processes, such

as mortality. More complex techniques consider both entrances (such as fertility) and exits in combination to estimate measures of reproductivity, intrinsic rates of growth or decline, and the parameters of stationary and stable populations. Techniques such as direct standardization or the calculation of shares or indexes of dissimilarity allow comparisons across two or more populations.

DATA SOURCES

The use of demographic methods to describe human populations is usually based on secondary data generated by national censuses; surveys; vital statistics registration systems; and, less frequently, population registers. The goal of a national CENSUS is to elicit a complete count of all people on a specific date, either all people present in the country (a *de facto* count), or all people residing in the country (a *de jure* count). In practice, census counts are marred by some degree of underenumeration or undercounting of some individuals (often assessed through CAPTURE-RECAPTURE models) and overcounting of other individuals. Most national censuses also include information such as the age, sex, relationship to household head, educational attainment, and marital status of each individual counted in the census.

Vital statistics registration systems describe the incidence of births, deaths, and sometimes other demographic events, such as marriages and divorces, as they occur over time. Vital statistics systems also often include additional information, such as age, sex, and marital status of the person experiencing the demographic event. Population registers combine the benefits of regular censuses and vital registration systems by keeping continuous track of the entire population and vital events, but they are difficult and expensive to maintain and thus much rarer.

BASIC DEMOGRAPHIC MEASURES

Proportions (or percentages) and ratios can be used to describe the structure of a population or the distribution of its elements along a variable of interest. Because many of the demographic and social processes producing change in a population are strongly age-graded, more complex demographic techniques often control for, or take into account, the age structure of the population. The complete age structure of a population can be described by tabulating the absolute

numbers of people or by calculating the proportions (or percentages) of people in each age interval. The results are often portrayed graphically in a POPULATION PYRAMID. It is also common to calculate summary measures describing a particular feature of the age-sex composition of a population. One important summary measure is the age-dependency ratio, defined as the sum of the number of people aged less than 18 and the number of people aged 65 and over (the “dependent population”) divided by the number of people aged 18–64 (the “working population”). Another summary measure is the sex ratio, defined as the number of males divided by the number of females.

MEASURES OF CHANGE

Almost all demographic methods and techniques describing change in a population build upon the basic demographic model in which the size of population at time t is a function of the size of the population at an earlier point in time, $t - n$, and the numbers of entrances and exits between time t and time $t - n$. A national population, for example, changes between time t and time $t + n$ through the quartet of classic demographic events: births (B), deaths (D), arrivals of in-migrants (I), and departures of out-migrants (O).

The most basic demographic measure describing change is the rate. A rate is the ratio of the number of demographic events occurring within a specified time frame (typically 1 year) to the number of people at risk of experiencing that event. The crude death rate for a population, for example, can be calculated as the total number of deaths during a specified year divided by the estimate of the total population midway through the year. Rates are often multiplied by a constant, k , such as 1,000 or 100,000. The crude death rate (CDR) of the United States in the year 2000, for example, was 8.5 deaths per 1,000 people:

$$\text{CDR} = \frac{D}{N} \times k = \frac{2,403,000}{281,422,000} \times 1,000 = 8.5.$$

The crude birth rate (CBR) of the United States in the year 2000, defined in parallel fashion to the CDR above (but with the number of births substituted for the number of deaths), was 14.4 births per 1,000 people.

$$\text{CDR} = \frac{B}{N} \times k = \frac{4,065,000}{281,422,000} \times 1,000 = 14.4.$$

The difference between the CBR and the CDR is the crude rate of natural increase (CRNI). In the case

Table 1 Measures of Reproductivity for the U.S. Population, 2000

<i>Age Interval</i>	<i>Age-Specific Fertility Rate Per 1,000 Women</i>	<i>Age-Specific Fertility Rate × Proportion of Births That Are Female</i>	<i>Proportion of Daughters Surviving to the Age of Their Mother</i>	<i>Product</i>
15–19	49.4	24.1	0.9902	23.9
20–24	112.3	54.8	0.9879	54.1
25–29	121.4	59.2	0.9854	58.4
30–34	94.1	45.9	0.9822	45.1
35–39	40.4	19.7	0.9776	19.3
40–44	8.4	4.1	0.9707	4.0
Sum	426.0	207.9		204.8
Sum × 5	2,130.0	1,039.5		1,023.8

SOURCES: Martin, Hamilton, Ventura, Menacker, and Park (2002); National Center for Health Statistics (2003).

NOTE: Age-specific fertility rates for the youngest and oldest age intervals include births to women younger than 15 and births to women older than 44, respectively. The age-specific fertility rates referring to female babies were estimated using a sex ratio at birth of 1,049 boys per 1,000 girls.

of the United States in 2000, the annual rate of natural increase was 5.9 people per 1,000.

$$\text{CRNI} = \text{CBR} - \text{CDR} = 14.4 - 8.5 = 5.9.$$

The crude in-migration rate or the crude out-migration rate can be calculated in a manner directly analogous to the crude birth rate or crude death rate. Many countries do not, however, collect comprehensive data on out-migration, and data on in-migration are often flawed. Thus, migration rates often refer to net migration ($I - O$).

DEATH RATES

More refined rates refer to subpopulations defined through age, sex, or a social characteristic, such as race or ethnicity. Because the force of mortality varies so strongly by age, death rates are often presented for specific age groups. The formula written below for an age-specific death rate uses a common style of notation in which the capital letters M , D , and N refer to the death (or mortality) rate, number of deaths, and number of people, respectively, and the right-hand and left-hand subscripts identify the lower bound and the width of the age interval, respectively. The death rate for people aged 15–24 in the United States in the year 2000 was 79.9 deaths per 100,000 young adults aged 15 to 24:

$${}_{10}M_{15} = \frac{{}_{10}D_{15}}{{}_{10}N_{15}} \times k = \frac{31,300}{39,183,900} \times 100,000 = 79.9.$$

If the focus is on infants, the infant mortality rate (IMR) is often substituted for the age-specific death rate for the population aged 0 to 1 because of common

errors in the census enumeration of infants. The infant mortality rate is calculated as the number of deaths to infants divided by the number of births. In the United States in 2000, preliminary data on infant deaths and births yield an infant mortality rate of 6.9 infant deaths per 1,000 births.

FERTILITY RATES

Fertility rates typically refer to women in the reproductive ages because of the difficulties gathering accurate data on the ages (or other attributes) of male parents. The formula below for an age-specific fertility rate uses the capital letters F , B , and N to refer to the fertility rate, the number of births born to women in the age interval, and the number of women in the age interval, with the subscripts referring, as before, to the lower bound and width of the age interval. The fertility rate for women aged 15–19 in the United States in the year 2000 was 48.5 births per 1,000 women:

$${}_5F_{15} = \frac{{}_5B_{15}}{{}_5N_{15}} \times k = \frac{469,000}{9,670,000} \times 1,000 = 48.5.$$

Summing the age-specific fertility rates within the reproductive age span of women (typically 15–44 or 15–49) and multiplying by the width of the age categories yields the total fertility rate (TFR). The sum of the age-specific fertility rates presented in column 1 of Table 1 shows that the total fertility rate for the United States in 2000 was 2,130 births per 1,000 women. The TFR can be interpreted as the total number of births to a hypothetical cohort of women experiencing the age-specific fertility rates observed in 2000 as they live through their reproductive life span.

MEASURES OF REPRODUCTIVITY

Classic measures of reproductivity focus on one sex, typically the female population. The calculation and interpretation of the gross reproduction rate (GRR) parallel the calculation and interpretation of the total fertility rate with one exception—the summed age-specific fertility rates consider only female births rather than all births and are sometimes called “maternity rates” rather than fertility rates. If age-specific maternity rates are unavailable, they are often estimated by multiplying age-specific fertility rates by the proportion of all births that are female. In the formula below, the right-hand superscript refers to the sex of the births:

$$\begin{aligned} \text{GRR} &= n \sum_n F_x^f = n \sum_n F_x \times \frac{B^f}{B} \\ &= 5(49.4 + 112.3 + \dots + 8.4) \\ &\quad \times \frac{1,000}{2,049} = 1039.5. \end{aligned}$$

The net reproduction rate (NRR) takes mortality into consideration by multiplying the age-specific maternity rates by the proportion of newly born daughters surviving to the age of their mother, ${}_nL_x/{}_nl_0$, which is derived from a LIFE TABLE for females. The results, displayed in detail in Table 1, show an NRR of 1,023.8 for the United States in the year 2000:

$$\begin{aligned} \text{NRR} &= n \sum_n F_x^f \times \frac{{}_nL_x^f}{{}_nl_0} \\ &= 5[(24.1 \times 0.9902) + (54.8 \times 0.9879) + \dots] \\ &= 1023.8. \end{aligned}$$

Thus, a cohort of 1,000 U.S. women subject to the age-specific fertility rates observed in the year 2000 could expect to have 2,130.0 births by the end of their reproductive life span. Of these, 1,039.5 would be daughters, and 1,023.8 would survive to the age of their mother.

Every population can be described at a moment in time by a schedule of age-specific fertility and mortality rates. If these rates were to continue indefinitely, the result would be a stable population marked by a birth rate, a death rate, a growth rate, and an age structure that are “intrinsic” to the schedules of age-specific vital rates and that do not depend on the age structure of the original population.

STANDARDIZATION

Because fertility and mortality rates differ greatly across age intervals, it can be misleading to compare

crude fertility and mortality rates across populations if the age structures of the populations differ. The techniques of direct and indirect standardization allow more appropriate comparisons of rates across populations, although the former approach is preferable to the latter (see Shryock, Siegel, & Associates, 1973).

A crude rate for a population can be rewritten as the sum of the product of age-specific rates and the age-specific proportions of people in each age interval. For example, the crude death rate for population i can be written as

$$\text{CDR}_i = \frac{{}_\omega D_{0i}}{{}_\omega N_{0i}} = \sum \left(\frac{{}_n D_{xi}}{{}_n N_{xi}} \right) \left(\frac{{}_n N_{xi}}{{}_\omega N_{0i}} \right).$$

A directly standardized rate (DSDR _{j}) for population j is calculated by multiplying the age-specific rates of population j by the age-specific proportions of the standard or reference population, population i :

$$\text{DSDR}_j = \sum \left(\frac{{}_n D_{xj}}{{}_n N_{xj}} \right) \left(\frac{{}_n N_{xi}}{{}_\omega N_{0i}} \right).$$

The directly standardized death rate can then be compared to the crude death rate of the standard population because they were calculated using the same age structure. When comparing the rates of two populations, the choice of the “standard” population is largely arbitrary, although it is generally better to use the one with the more regular age structure. However, conventions sometimes govern the choice of the standard population. The U.S. Census Bureau, for example, recently declared the 2000 U.S. population to be the standard for historical comparisons involving the American population.

The difference between any crude rate and a directly standardized rate ($\text{CDR}_i - \text{DSDR}_j$) is the sum of the differences in the age-specific rates weighted by the age composition of the standard population. It is also possible to calculate the corresponding directly standardized rate for population j . The raw difference between the crude death rates of the two populations ($\text{CDR}_i - \text{CDR}_j$) can then be decomposed into three parts: a component attributable to differences in rates ($\text{CDR}_i - \text{DSDR}_j$); a component attributable to differences in age structure ($\text{CDR}_i - \text{DSDR}_i$); and an interaction term ($\text{DSDR}_j + \text{DSDR}_i - \text{CDR}_i - \text{CDR}_j$). Some scholars divide the interaction term between the rate and age components (see Preston, Hueveline, & Guillot, 2001).

Direct standardization and its corollary, the decomposition of the difference between two rates (or means), is a technique with wide applicability outside of the

classic demographic concerns of fertility or mortality. It can be used whenever a rate (or mean) can be conceptualized as the sum of the product of i -specific rates and the relative proportion of the population along dimension i . It is also possible to standardize rates (and to decompose differences between rates) along two or more dimensions of interest (see RATE STANDARDIZATION).

MEASURES OF INEQUALITY AND SEGREGATION

Some demographic measures and techniques assess and compare the distribution of counts across units of analysis. For example, income (a count of number of dollars earned) may be distributed unequally across households (the unit of analysis) in a population. The distribution of such variables is often described through “shares.” Shares are calculated by first ranking the units of analysis by the values of the count variable, from lowest to highest, and dividing the units of analysis into groupings of equal size, typically quintiles (fifths) or deciles (tenths). In 2000, for example, the poorest fifth of all U.S. households earned 3.6% of the grand total of all income earned by U.S. households, whereas the top quintile earned 49.7%.

Alternatively, a population (counts of people) can be distributed unequally across the categories of a variable (e.g., geographic units). The index of dissimilarity (D) is often used to compare the distributions of two or more populations across a categorical variable. The index of dissimilarity is calculated using the formula

$$D = \frac{1}{2} \sum_i \left(\left| \frac{y_i}{\sum_i y_i} - \frac{x_i}{\sum_i x_i} \right| \right).$$

For example, the categorical variable could refer to the four major regions in the United States (Northeast, Midwest, South, and West); y to African Americans; and x to Whites. The numbers (in 1,000s) of African Americans living in each of the four regions in the year 2000 are 6,100; 6,500; 18,982; and 3,077; the corresponding numbers for Whites are 41,534; 53,834; 72,819; and 43,274. A value of 0 for the index of dissimilarity would suggest that African Americans and white Americans are distributed similarly across the four regions, whereas a value of 1 would suggest that the two populations live in entirely different regions. The value of the index of dissimilarity is .20, suggesting that about a fifth of either African Americans or Whites

would have to shift to another region for the two distributions to agree exactly. Although the index of dissimilarity is easy to calculate and interpret, it has several shortcomings. It is, for example, sensitive to the number of categories. Other indexes, such as the GINI COEFFICIENT (or Gini ratio) or the entropy index (H), can be calculated in its stead, although they, too, have shortcomings. See Siegel (2002) for a discussion of these alternatives.

—Gillian Stevens

REFERENCES

- Martin, J. A., Hamilton, B. E., Ventura, S. J., Menacker, F., & Park, M. M. (2002, February 12). *Births: Final data for 2000* (National Vital Statistics Report, Vol. 50, No. 5). Hyattsville, MD: National Center for Health Statistics.
- National Center for Health Statistics. (2003). *United States life tables, 2000*. Retrieved from www.cdc.gov/nchs/data/lt2000.pdf.
- Preston, S. H., Heuveline, P., & Guillot, M. (2001). *Demography: Measuring and modeling population processes*. Oxford, UK: Basil Blackwell.
- Shryock, H. S., Siegel, J. S., & Associates. (1973). *The methods and materials of demography*. New York: Academic Press. (Condensed edition by Edward G. Stockwell.)
- Siegel, J. S. (2002). *Applied demography*. San Diego, CA: Academic Press.

DEPENDENT INTERVIEWING

Dependent INTERVIEWING is a technique used in some LONGITUDINAL RESEARCH such as PANEL surveys, where the same individuals are reinterviewed at subsequent points in time. Depending on the survey design, interviews may take place quarterly, annually, biannually, or over some other specified period. Dependent interviewing can be defined as a process in which a question asked of a respondent within the current survey wave is informed by data reported by a respondent in a previous wave (Mathiowetz & McGonagle, 2000). For example, if, at one survey interview, a respondent reported being employed as a bus driver, dependent interviewing would use that response in the question asked about the respondent's job at the following interview. Instead of simply asking the respondent “What is your current job?” at both interview points, a dependent interviewing approach might ask, “Last year you told us you were working as a bus driver. Is that still

the case?" The respondent can then either confirm the previous response or give a different answer if some change has occurred.

Longitudinal surveys are subject to the types of error normally found in survey data, including recall error, interviewer effects, and "noise" in the coding process. Longitudinal data, where continuous records of activities such as labor market histories or continuous income records are collected, suffer from an additional measurement problem known as the *seam effect* or the *seam problem* (Doyle, Martin, & Moore, 2000; Lemaitre, 1992). The seam effect occurs when an artificially high level of observed change in activity spells at the seam between two survey periods is reported by respondents. Respondents tend to report events that they have already reported in the previous interview and pull them into the current interview period, known as *forward telescoping*, with the result that events become bunched around the seam between the two survey periods. The use of dependent interviewing is designed to reduce these effects and produce more consistent data over time. Dependent interviewing has become more common with the advent of computer-assisted interviewing technologies for data collection, which have made it technically feasible to feed forward previously reported information from one interview into the next interview. Dependent interviewing has benefits in terms of data quality, reducing respondent and interviewer burden, as well as the cost of data collection and processing. Despite this, there are also concerns. Depending on how it is implemented, dependent interviewing could result in an underestimate of the true level of change within the population because it is easier for the respondent to simply agree with his or her previous report.

—Heather Laurie and Mike C. Merrett

REFERENCES

- Doyle, P., Martin, B., & Moore, J. (2000). *Improving income measurement*. The Survey of Income Program Participation (SIPP) Methods Panel. Washington, DC: U.S. Bureau of the Census.
- Lemaitre, G. (1992). *Dealing with the seam problem* (Survey of Labour and Income Dynamics Research Papers 92-05). Ottawa: Statistics Canada.
- Mathiowetz, N., & McGonagle, K. (2000). An assessment of the current state of dependent interviewing in household surveys. *Journal of Official Statistics*, 16(4), 401-418.

DEPENDENT OBSERVATIONS

Dependent observations are observations that are somehow linked or clustered; they are observations that have a systematic relationship to each other. Clustering typically results from research and sampling design strategies, and the clustering can occur within space, across time, or both.

Single and multistage cluster samples produce samples that are clustered in space, typically geographical space. Identifying and selecting known and defined clusters, and then selecting observations from the chosen clusters, produce observations that are not independent of each other. Rather, the observations are connected by the cluster, whether it is a geographic entity (e.g., city), organization (e.g., school), or household.

Longitudinal research designs produce observations that are linked across time. For example, even if individuals are selected randomly, a researcher may obtain an observation (e.g., attitude_{it}) on each individual i at various points in time, t . The resultant observations (e.g., attitude_{i1} and attitude_{i2}) are not independent because they are both attached to person i .

The occurrence of dependent observations violates a key assumption of the general linear model: that the error terms are independent and identically distributed. Observations that are linked in space or across time are typically more homogeneous than independent observations. Because linked observations do not provide as much "new" information as independently selected observations might, the resultant sample characteristics are less heterogeneous. In QUALITATIVE RESEARCH, this means that researchers should make inferences cautiously (King, Keohane, & Verba, 1994). In quantitative research, this means that standard errors of estimates will be deflated (Kish, 1965).

To correct for dependent observations resulting from cluster sampling techniques, sampling weights should be used if they are a function of the dependent variable, and thus the error term. In addition, Winship and Radbill (1994) recommend using the White heteroskedastic consistent estimator for standard errors instead of trying to explicitly model the structure of the error variances. This estimator can also be used to correct for multiplicity in the sample and is available in standard statistical packages.

To correct for temporally dependent observations, researchers typically assess and model the order of

the temporal autocorrelation and introduce lagged endogenous and/or exogenous variables. Such techniques have been studied extensively in the time-series literature.

Instead of simply correcting for dependence, researchers are increasingly interested in modeling and explaining the nature of the dependence. Links between individuals or organizations, and clusters of such units, are of primary interest to social science researchers. For decades, network theorists and analysts have focused their efforts on describing and explaining the nature of dependence observed within a cluster or clusters; recent advancements extend these methods to include affiliation networks (Skvoretz & Faust, 1999) and methods for studying networks over time (Snijders, 2001). More recently, Bryk and Raudenbush (1992) have developed a framework for modeling clustered, or what they call “nested,” data. For such HIERARCHICAL (NON)LINEAR MODELS (also known as mixed models, because they typically contain both fixed and random effects), cross-level interactions in MULTILEVEL ANALYSIS permit an investigation of contextual effects: how higher-level units (schools, organizations) can influence lower-level relationships.

—Erin Leahey

REFERENCES

- Bryk, A., & Raudenbush, S. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.
- King, G., Keohane, R. D., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.
- Kish, L. (1965). *Survey sampling*. New York: Wiley.
- Skvoretz, J., & Faust, K. (1999). Logit models for affiliation networks. *Sociological Methodology*, 29, 253–280.
- Snijders, T. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, 31, 361–395.
- Winship, C., & Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research*, 23(2), 230–257.

DEPENDENT VARIABLE

A dependent variable is any variable that is held to be causally affected by another variable, referred to as the INDEPENDENT VARIABLE. The notion of what can be taken to be a dependent variable varies between experimental and nonexperimental

research [see entries on DEPENDENT VARIABLE (IN EXPERIMENTAL RESEARCH) and DEPENDENT VARIABLE (IN NONEXPERIMENTAL RESEARCH)].

—Alan Bryman

DEPENDENT VARIABLE (IN EXPERIMENTAL RESEARCH)

A dependent variable (DV) is a measurement. In well-designed experimental research, the DV is the measure of the effect of the INDEPENDENT VARIABLE (IV) on participants' responses. The measure is termed dependent because it is expected to be influenced by (or is dependent on) systematic variations in the independent variable. With appropriate experimental design, observed variations in the DV are attributable to the effects of variation of the independent variable, and hence are fundamental to the attribution of cause.

Strictly speaking, in nonexperimental research, all variables are considered dependent, insofar as they are not systematically manipulated. Thus, in correlational studies, all variables are dependent, because such research uses no manipulation. In multiple regression prediction research, dependent measures are used as predictors and sometimes are termed independent (or predictor) variables. In some research traditions, predictors are sometimes interpreted causally if they are clearly unlikely to serve as outcomes. For example, a series of naturalistic studies of ambient temperature and aggression suggests a moderately strong positive relationship. It is extremely unlikely that people's aggression affects the temperature, so the correlation is interpreted as heat causes people to become more aggressive. Relationships of this sort, based on common sense or theory, often are interpreted causally, but the inferences made are tentative and not as clear as those drawn from experimental contexts, because extraneous variables that may co-occur with the critical variable may prove to be the true causal factor. For instance, in the heat-aggression example, it may be that the tendency of people to go out on the streets in hot weather to escape the heat of their homes is the true instigating factor for aggression. Crowds (of unhappy people) may be the necessary ingredient in the equation.

The DV may take on many forms. The most highly self-conscious form of DV is the self-report in which

participants express an attitude, belief, or feeling; indicate their knowledge about a phenomenon; make a causal attribution about their, or another's, actions; and so on. Measures of this type must be interpreted with caution for at least two reasons. First, people generally are reluctant to report socially undesirable actions or attitudes. A DV used to assess offensive acts or beliefs might evoke socially desirable, but untrue, responses. A more pernicious problem arises from the tendency of participants to attempt to assist the investigator by answering in such a way as to help confirm the research hypotheses. Overcooperative responding to the perceived needs of the experimenter (see DEMAND CHARACTERISTICS) can prove misleading and result in a confirmation of an incorrect hypothesis.

At the opposite end of the self-conscious response continuum are DVs collected on respondents who are unaware they are being observed. Such unobtrusive or nonreactive measures are common in research conducted outside formal laboratory settings (see LABORATORY EXPERIMENT) and offer the advantage of truthful, or at least non-self-conscious, response on the part of unwary participants. However, such measures can prove insensitive, and often are less reliable than more straightforward self-report measures (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). In addition, the ethical justifiability of such measures has come under intense debate. As such, they should be used only after very careful consideration.

In addition to the self-conscious/non-self-conscious responding dimension, DVs also differ in terms of structure. Close-ended DVs are highly structured; they may involve true/false or multiple-choice response formats. These forms of DV are most appropriate in hypothesis-testing research, where the issues under investigation are precisely specified and the requisite information clearly specified on the basis of theory. In less formal, hypothesis-generating research, such restrictive measures may prove unsatisfactory, and more open-ended measures are recommended. Such measures may involve the researcher's field notes of an observational setting, notes on an unstructured conversation, and so on. In these instances, the format of the observation is open. Although open-ended DVs require coding and theory-based interpretation to support the research process, they may supply useful insights that may be developed to support more focused research efforts.

—William D. Crano

REFERENCE

- Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. (1981). *Nonreactive measures in the social sciences*. Boston: Houghton-Mifflin.

DEPENDENT VARIABLE (IN NONEXPERIMENTAL RESEARCH)

A dependent variable is held to be causally affected by an INDEPENDENT VARIABLE. Strictly speaking, a dependent variable is randomly distributed, although this randomness can be influenced by a number of structural factors and in the most rigorous research setting by an EXPERIMENT in which the independent variable is manipulated. However, in nonexperimental research, variables are not manipulated, so that it is sometimes deemed to be unclear which among the various variables studied can be regarded as DEPENDENT VARIABLES. Sometimes, researchers use the term loosely to refer to any variables they include on the left-hand side of a REGRESSION equation where causal inference is often implied. But it is necessary to discuss what qualifies as a dependent variable in different RESEARCH DESIGNS. When data derive from a CROSS-SECTIONAL DESIGN, information on all variables is collected coterminously and no variables are manipulated. This means that the issue of which variables have causal impacts on other variables may not be immediately obvious. Although it is difficult to establish causality, especially when using cross-sectional data, researchers use a few rules of thumb as to which variable to regard as dependent. Five such rules, which are not necessarily mutually exclusive, may serve as a guide for use in both cross-sectional and longitudinal research.

1. One variable precedes another in time. Even in cross-sectional surveys, certain variables record events that occurred earlier than others, such as the educational attainment of the respondent's parent versus the educational attainment of the respondent, which can be regarded as dependent.

With LONGITUDINAL RESEARCH, such as a PANEL design, the issue is somewhat more complex, because although no independent variables are manipulated, the fact that data can be collected at different points in time allows some empirical leverage on the issue of the temporal and, hence, causal priority of variables. Strict experimentalists might still argue that the lack of

manipulation casts an element of doubt over the issue of causal priority. To deal with this difficulty, analysts of panel data resort to the so-called cross-lagged panel analysis for assessing causal priority among a pair of variables measured at two or more points in time. The line between dependent and independent variables in such analysis is, in a sense, blurred because the variable considered causally precedent is treated as an independent variable in the first regression equation but as a dependent variable in the second. Conversely, the variable considered causally succedent is treated as dependent in the first equation and as independent in the second. Judging from analyses like this, it appears that the term “dependent variable” can be quite arbitrary; what matters is clear thinking and careful analysis in nonexperimental research.

2. Variables are from a well-known sequence. For example, events in one’s LIFE COURSE, such as first sexual intercourse and first birth, have a clear causal order.

3. One event freezes before another starts. The amount of donation to one’s alma mater may depend on the type of bachelor’s degree one receives. The act of donation does not begin until the degree is completed.

4. One variable is more changeable than another. To take a simple example: Imagine that we conduct a survey that shows that gender and a scale of political conservatism are related. One’s gender hardly changes, but one’s political orientation may. It seems reasonable to assume that political conservatism is the dependent variable. Of course, there may be INTERVENING VARIABLES or MODERATING VARIABLES at work that influence the causal path between gender and conservatism, but the basic point remains that it is still possible and sensible to make causal inferences about such variables.

5. One variable is more fertile than another. The more fertile (i.e., more likely to produce consequences) one, say, one’s religious affiliation, could be used as an independent variable, whereas the less fertile one, say, the number of children one has, would be the dependent variable.

Researchers employing nonexperimental research designs (such as a SURVEY design) must engage in *causal inference* to tease out which are the dependent variables. Essentially, this process entails a mixture of commonsense inferences from our understanding about the nature of the social world, including the five rules discussed earlier, and from existing theory and research, as well as analytical methods in the

area that is the focus of attention. It is this process of causal inference that lies behind such approaches as CAUSAL MODELING and PATH ANALYSIS to which the cross-lagged panel analysis is related.

Making causal inferences becomes more difficult when causal priority is less obvious. For example, if we find, as a result of our investigations in a large firm, that employees’ levels of routine in their jobs and their organizational citizenship behavior are negatively correlated, it may be difficult to establish direction of causality, although existing findings and theory relating to these variables might provide us with helpful clues. It could be that employees in more routine jobs are less inclined to put themselves out for the firm, but it could also be the case that employees who exhibit organizational citizenship behavior are less likely to be assigned to routine jobs. A panel study is often recommended in such circumstances as a means of disentangling cause and effect.

—Alan Bryman and Tim Futing Liao

REFERENCES

- Davis, J. A. (1985). *The logic of causal order* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–055). Beverly Hills, CA: Sage.
- Lieberson, S. (1987). *Making it count: The improvement of social research and theory*. Berkeley: University of California Press.
- Rosenberg, M. (1968). *The logic of survey analysis*. New York: Basic Books.

DESCRIPTIVE STATISTICS. See UNIVARIATE ANALYSIS

DESIGN EFFECTS

The design effect indicates the impact of the sample design on the VARIANCE of an estimate. The impact is measured relative to the variance of the equivalent estimate obtained from a SIMPLE RANDOM SAMPLING of the same size. Thus, the design effect for an estimate of a POPULATION parameter y can be written

$$\text{DEFF}(\hat{y}_{\text{Com}}) = \frac{V(\hat{y}_{\text{Com}})}{V(\hat{y}_{\text{SRS}})}, \quad (1)$$

where $\hat{y}_{\text{Com}}$ is the estimate of y under the complex sample design in question and $\hat{y}_{\text{SRS}}$ is the estimate from a simple random sample of the same size.

The design effect is specific to both the estimate and the sample design. Consequently, estimates of different quantities from the same survey may have different design effects, as may estimates of the same quantity from different SURVEYS. Three important influences on the design effect are sample clustering, sample stratification, and sample inclusion probabilities.

Multistage clustered sample designs (see MULTISTAGE SAMPLING) tend to result in design effects greater than 1. This is because sample units within clustering units tend to be more similar to one another than are sample units in the population as a whole (i.e., clusters are relatively homogeneous). This may be true, for example, of households living within small areas, pupils attending the same school, employees of the same employer, and so on. The design effect due to clustering will depend on the extent of the homogeneity within clusters and the (mean) cluster sample size:

$$\text{DEFF}(\hat{y})_{\text{Clus}} = 1 + (b - 1)\rho, \quad (2)$$

where b is the (mean) cluster sample size and ρ is the intracluster CORRELATION coefficient.

The design effect due to proportionate stratification (see STRATIFIED SAMPLE) will tend to be less than 1. Again, this is because strata are relatively homogeneous. But unlike clustering, which limits the sample to a subset of clusters, stratification ensures that the sample represents all strata. The higher the correlation between stratum and target variable, the smaller (more beneficial) the design effect due to stratification.

The design effect due to variable inclusion probabilities (encompassing both selection probabilities and response probabilities) will tend to be greater than 1 and will tend to be greater the greater the variation in the probabilities. There are notable exceptions, such as the use of optimal allocation to strata in order to increase precision of estimation in the situation where stratum population variances are known or can be estimated. The design effect due to variable inclusion probabilities can be written as follows:

$$\text{DEFF}(\hat{y})_{\text{Inc}} = \frac{n}{N^2 S^2} \sum_{h=1}^H \frac{N_h^2 S_h^2}{n_h}, \quad (3)$$

where n_h , N_h , and S_h^2 are, respectively, the sample size, population size, and population variance in class h , the

population being divided into H mutually exclusive and comprehensive classes, whereas n , N , and S^2 are the overall sample size, population size, and population variance.

Many survey technical reports publish tables of design effects for a range of estimates. These can be useful in informing the design of future surveys as well as the interpretation of results from the current survey.

A closely related concept is the *design factor*, which indicates the impact of the sample design on the STANDARD ERROR of the estimate, and hence on the width of the CONFIDENCE INTERVAL:

$$\begin{aligned} \text{DEFT}(\hat{y}_{\text{Com}}) &= \frac{s.e.(\hat{y}_{\text{Com}})}{s.e.(\hat{y}_{\text{SRS}})} \\ &= \sqrt{\text{DEFF}(\hat{y}_{\text{Com}})}. \end{aligned} \quad (4)$$

—Peter Lynn

REFERENCES

- Kish, L. (1965). *Survey sampling*. New York: Wiley.
Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11(1), 55–77.

DETERMINANT

The determinant of a square MATRIX is a unique scalar associated with that matrix. The determinant of a matrix $\mathbf{A}$ is most commonly denoted as $\det \mathbf{A}$ or $|\mathbf{A}|$. The notation for determinants substitutes vertical lines for the brackets of matrices. Thus, for

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad |\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}.$$

Texts may sometimes refer to

$$|\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

as the determinate function and the resulting scalar as the value of the determinant.

Calculation of the determinant is straightforward for matrices of order 1 and 2. If $\mathbf{A} = [a_{11}]$, then $|\mathbf{A}| = |a_{11}| = a_{11}$. If

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

then

$$|\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Note that the determinant for the second-order example is the product of the elements on the main diagonal minus the product of the elements on the off-diagonal. A technique known as row (or column) expansion allows for the systematic calculation of determinants for matrices of order 3 and above. In brief, for matrix $\mathbf{A}$ of order n , row expansion produces the equation

$$|\mathbf{A}| = \sum_{j=1}^n (-1)^{i+j} a_{ij} |\mathbf{A}_{ij}|$$

for any one row $i = 1, 2, \dots, n$. For column expansion, the equation is

$$|\mathbf{A}| = \sum_{i=1}^n (-1)^{i+j} a_{ij} |\mathbf{A}_{ij}|$$

for any one column $j = 1, 2, \dots, n$. In these equations, $|\mathbf{A}_{ij}|$ is the determinant of the submatrix formed by eliminating the i th row and j th column of $\mathbf{A}$. This is often referred to as the minor of a_{ij} from $\mathbf{A}$. Thus, expanding by the first row for matrix $\mathbf{A}$ of order 3,

$$|\mathbf{A}| = \sum_{j=1}^3 (-1)^{1+j} a_{1j} |\mathbf{A}_{1j}| = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}.$$

Several theorems allow the manipulation of the rows and columns of a determinant to facilitate the calculations. Among them are the following:

- The determinant of a matrix $\mathbf{A}$ is equal to the determinant of the transpose of $\mathbf{A}$ (i.e., $|\mathbf{A}| = |\mathbf{A}^T|$).
- Exchanging any two rows of a determinant changes its sign.
- If every element of a row is multiplied by a factor k , then the value of the determinant is multiplied by k .
- The value of a determinant is unchanged if a scalar multiple of one row is added to another row.

If the determinant of a matrix is nonzero, the matrix is said to be nonsingular and will have an

inverse. Determinants are used to calculate the inverses of matrices, both of which may then be used to solve systems of equations such as those found in REGRESSION analysis. Determinants are also used with Cramer's rule, a method of solving SIMULTANEOUS linear EQUATIONS, and in determining whether the extreme points of FUNCTIONS are maxima or minima.

If the determinant of a matrix equals zero, the matrix is said to be singular. This indicates that two or more of the rows or columns are LINEARLY DEPENDENT. Algebraically, this means that an inverse of the matrix cannot be calculated. As a practical matter, it means that two or more of the variables in the model represented by the matrix are linear TRANSFORMATIONS of each other.

—Timothy M. Hagle

REFERENCES

- Cullen, C. G. (1990). *Matrices and linear transformations* (2nd ed.). New York: Dover.
- Pettoufrezza, A. J. (1978). *Matrices and transformations*. New York: Dover.

DETERMINATION, COEFFICIENT OF.

See R-SQUARED

DETERMINISM

Determinism is a term usually used pejoratively to describe an argument or methodology that simplistically reduces causality to a single set of factors acting more or less directly to produce outcomes. So, for example, economic determinism attributes social, political, and cultural effects to economic relations as the fundamental causal factors; technological determinism does the same with technological innovation; textual determinism attributes the impact of mediated messages on individuals and audiences to the textual attributes of the messages or the medium, and so on. A related but different term is reductionism, where an analysis is accused of reducing a complex set of circumstances to a limited set of supposedly salient characteristics.

The methodologies most often accused of determinism are those that favor a totalizing approach

to explanation, such as Marxism, Freudianism, informationalism, textualism, systems theory, and so on; those that want to privilege certain natural, social, or physical characteristics as fundamental to social activity, such as climate, class, gender, race, or geography; or those that single out particular spheres of human activity as overwhelmingly powerful in their impact on other spheres, such as technology, sex, or work.

Of course, there is poor intellectual work that manifests determinist and/or reductionist flaws. But what often underlies debates about determinism are the more fundamental issues of where a researcher might look for adequate causal explanation in human phenomena, whether such a search is worth undertaking at all, whether fields of human activity in time and space can be thought of as totalities (or sets of overlapping totalities), and whether there are systematic hierarchies of factors in social causation. In short, arguments about determinism can be proxy arguments about causality.

Thinkers have approached causality in a number of different ways. One interesting outcome of sociometric modeling has been questions about quantifiable predictability as a corollary of causality: The debate is between deterministic, mathematical models and stochastic, statistical approaches (the latter incorporating chance outcomes). At the metatheoretical level, some argue that the structure of a whole determines the nature and function of its parts; systems theory approaches are typical of this approach. Others argue that the whole is the sum of its parts, and that only the parts are knowable in any worthwhile sense; pragmatic approaches often take this position.

A third approach looks to neither the parts nor the whole but to the relations between forces in constant conflict that structure the particular parts in relation to any whole at any specific juncture of time and space. This conflict produces outcomes in particular periods and places that determine the nature of relationships between the parts and so structure the whole. It is an approach that developed in the second half of the 20th century, in different disciplines and from different totalizing perspectives (e.g., the educational psychologist Jean Piaget called it “operational structuralism,” and for the Marxist geographer David Harvey, it is “dialectical, historical-geographical materialist theory”).

It is an approach that views causation as a complex and continuing process of conflict among abstract forces, manifesting themselves in specific material

ways in specific historical and geographical contexts, producing outcomes that themselves then feed into the process of further causation. It rejects notions of direct, linear causality between factors, but does seek to identify hierarchies of power or dominance in the complex relationships that determine outcomes.

David Harvey’s work is preoccupied with these methodological questions with respect to urbanization and presents a thorough exploration of the spatial and temporal dimensions of causality. *Social Justice and the City* (1973) is an early, detailed exposition of the methodological issues; *The Condition of Postmodernity* (1989) is his best known work.

—Chris Nash

REFERENCES

- Harvey, D. (1973). *Social justice and the city*. Baltimore: Johns Hopkins University Press.
 Harvey, D. (1989). *The condition of postmodernity*. Oxford, UK: Basil Blackwell.

DETERMINISTIC MODEL

A deterministic model is one that contains no random elements.

Deterministic models began to be widely used to study physical processes in the early 18th century with the development of differential equations by mathematicians such as Jakob Bernoulli, Johann Bernoulli, and Leonhard Euler. Differential equations, which give the value of a variable as a function of its value at an infinitesimally earlier time point, remain the most common form of deterministic MODEL. Differential equations proved applicable to a wide variety of astronomical and mechanical phenomena, and have produced a rich mathematical literature. Finite difference equations—where the value of a variable is a function of its value at a time period lagged by a finite value (typically 1)—are another form of deterministic model, albeit a form that has proven to be less mathematically elegant than differential equations. Difference equations have been commonly used to model biological and social processes where change occurs more slowly, for example, across generations in a biological population or across budget years in an economic process.

The values of a deterministic model are completely predictable provided one knows the functional form

of the model, the values of its coefficients, and the initial value of the process. During the 18th and 19th centuries, the physical universe was assumed by many natural scientists to operate according to purely deterministic laws. This meant, in theory at least, that it was entirely predictable—in the widely quoted formulation of the French mathematician Pierre Simon de Laplace in his work, *Analytic Theory of Probability* (1812):

Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective positions of the beings which compose it . . . it would embrace in the same formula both the movements of the largest bodies in the universe and those of the lightest atom; to it nothing would be uncertain, and the future as the past would be present to its eyes. (Mackay, 1977, p. 92)

As Laplace notes, a curious feature of deterministic models is that not only can the future be predicted from the model, but one can also deduce behavior in the past. Thus, for example, astronomers modeling the orbit of a newly discovered comet with deterministic equations can figure out not only where the comet is going, but where it came from. Similarly, a deterministic model of the path of a ballistic missile can be used to determine its approximate launch point, even if the launch was not observed.

The 20th century saw three challenges to the concept of a deterministic universe. First, early in the century, the theory of quantum mechanics developed and argued that at an atomic level, physical processes had intrinsically random elements, in contrast to the purely deterministic world envisioned earlier. Second, throughout the 20th century, the mathematical fields of probability and statistics—including the study of STOCHASTIC processes, which incorporate randomness—developed to a level of sophistication comparable to the 19th-century mathematics of deterministic processes. Subsequent work in the social sciences has consequently focused primarily on models with explicitly specified random elements.

Finally, near the end of the 20th century, chaotic processes began to receive considerable attention (see Kiel & Elliott, 1996). Chaotic models are deterministic—they are typically nonlinear difference equations—but generate sequences of values that appear to be random, and those values cannot be

predicted over long periods unless the initial values can be assessed with infinite precision, which is not possible in practice.

Despite these challenges, deterministic models remain useful in a variety of contexts. Notwithstanding quantum mechanics and chaos theory, deterministic equations can provide extremely accurate predictions for many physical processes that occur above the atomic level. Deterministic equations are also useful as approximations for social processes, even when these are known to contain random components. Examples include the growth of populations, budgets, and international arms races.

—Philip A. Schrodt

REFERENCES

- Kiel, L. D., & Elliott, E. (1996). *Chaos theory in the social sciences*. Ann Arbor: University of Michigan Press.
Mackay, A. L. (1977). *Scientific quotations*. New York: Crane, Russak.

DETRENDING

Often, TIME-SERIES data may exhibit an upward or downward trend over time. A trend may be present in the form of a change in the mean value and/or VARIANCE over time. A plot of a time series may be linear, demonstrating a trending mean over time. Regressing one such variable on another will often yield a high *R-SQUARED*, yet the estimated parameters may not reflect the relationship between the two variables. Rather, the estimated parameters will be BIASED upwards because of the trend, resulting in a SPURIOUS RELATIONSHIP between the dependent and independent variables, reflecting the common trend present in both of them. When this is the case, observations may be detrended before fitting a model in order to obtain true estimates of the relationship between the dependent and independent variables. Figure 1 shows a series that is trending over time, and that same series after the trend has been removed.

A simple solution to this problem is to include a time, or trend, variable as an INDEPENDENT VARIABLE in a REGRESSION analysis. The trend variable need not be linear; if the series were to exhibit a cubic trend, then the trend variable could be X^3 , where X is time. Furthermore, some time series exhibit seasonal

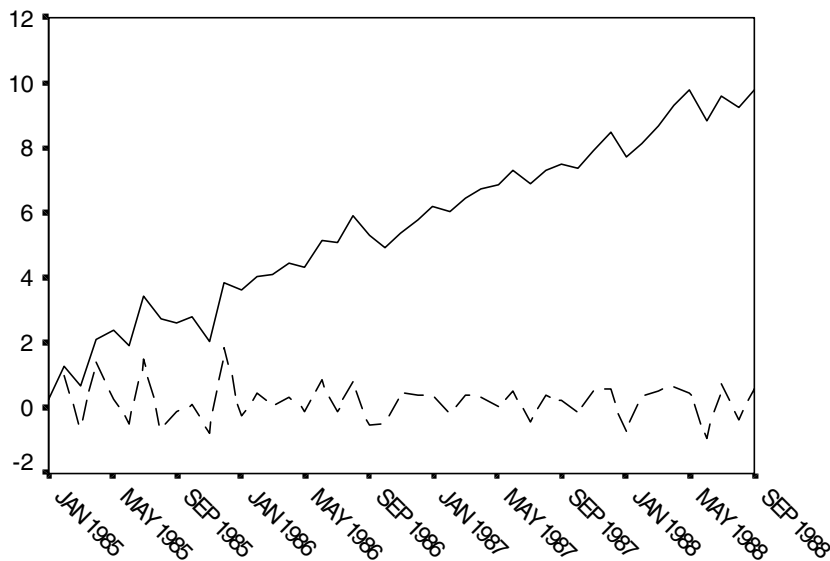


Figure 1 Example of a Trending Series and Same Series After Detrending

trends, and it is appropriate in these instances to include seasonal trend variables as well. Introducing a trend variable on the right-hand side of the equation will yield for its coefficient an estimate of the true relationship between it and the dependent variable.

If the dependent and independent variables exhibit different trends, say, the dependent a linear trend and the independent a quadratic, then it may be useful to employ a more complex detrending process. First, regress the dependent variable on time to obtain the RESIDUALS, $\hat{u}_{1t}$. Then, regress the independent variable on time to obtain the residuals, $\hat{u}_{2t}$. Finally, regress $\hat{u}_{1t}$ on $\hat{u}_{2t}$. With the time trend removed from both variables, the coefficient on $\hat{u}_{2t}$ will be the true relationship between the dependent and independent variable. This method for detrending is additionally useful if you wish to report not merely the relationship between the dependent variable and the independent variable, but their unique trends over time as well.

Detrending in this manner is only applicable for the series in question if the trend is deterministic. If the trend changes over time, it is said to be STOCHASTIC, and in this instance, ARIMA modeling may be an appropriate method to use to discover the underlying relationship. A UNIT ROOT test may be appropriate to determine whether the series is stationary. Trends are removed in ARIMA modeling by differencing series, or by subtracting previous values of the variable from

the value at time t . Simple linear trends, such as those easily removed by including a trend variable in regression, may also be removed in ARIMA models with a first-differenced equation, or an equation where the dependent variable is change in y from time, t , to previous time, $t - 1$.

—Andrew J. Civettini

REFERENCES

- Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). St. Louis, MO: McGraw-Hill.
- Harvey, A. C. (1990). *The econometric analysis of time series* (2nd ed.). New York: Philip Allan.

DEVIAANT CASE ANALYSIS

Deviant case analysis is a procedure used in CONVERSATION ANALYSIS (CA) for checking the validity and generality of proposed phenomena of conversational interaction. It bears comparison to hypothesis testing and falsification as general methodological principles. However, it has special features that make it an essential ingredient in CA.

The main aim of CA is to discover robust rules of conversational sequencing that are not merely statistically observed regularities, but ones that are demonstrably used and oriented to, as normatively expectable, by conversational participants. Through inductive analysis of a corpus of conversational data, potential rules of conversational order may be proposed, such as in greetings exchanges, or in the ways in which invitations may be either accepted or rejected. Deviant cases are those that do not appear to conform to the proposed rule or pattern.

The importance of these cases, beyond their status as disconfirmations, is that they are often the examples that turn out to best demonstrate participants' normative orientations, somewhat echoing the "breaching experiments" that were characteristic of early ETHNOMETHODOLOGY. Unlike those experiments, however, CA's deviant cases are ones found naturally occurring in a corpus of recorded talk.

Stephen Clayman and Douglas Maynard (1995) define three uses of deviant cases in CA: (a) Deviant cases may turn out, upon analysis, to instantiate the same normative orientations as are proposed for the standard cases; (b) a new analysis may be proposed that now takes full account of both the (apparently) deviant cases, and also the main set; and (c) a deviant case may turn out to exhibit a different kind of order in its own right, such that it properly belongs in a different collection.

The classic example of the second type, where a revised analysis is proposed, is an early CA study by Emanuel Schegloff (1968) on the ways in which telephone conversations routinely begin. Using a corpus of 500 examples, Schegloff had proposed a “distribution rule for first utterances,” to the effect that the person answering the call is, expectably, first to speak. One case was found to deviate from the rule. This led to a questioning and radical revision of what the rule proposed. Rather than proposing a rule in which call answerers make first turns, which might be considered first parts of a greeting exchange (e.g., “hello,” answered by “hello”), a revised rule could be proposed that accounted for all 500 cases. This was the summons-answer sequence, in which the first utterance of “hello” (or whatever) was now understood as the second part of a different kind of pair, the answer to the summons, which was the phone ringing.

—Derek Edwards

REFERENCES

- Clayman, S. E., & Maynard, D. (1995). Ethnomethodology and conversation analysis. In P. ten Have & G. Psathas (Eds.), *Situated order: Studies in the social organization of talk and embodied activities* (pp. 1–30). Washington, DC: University Press of America.
- Schegloff, E. A. (1968). Sequencing in conversational openings. *American Anthropologist*, 70, 1075–1095.

DEVIATION

Deviation is a very general term referring to some type of discrepancy. A simple example is the deviation or difference between an observation and the POPULATION mean, μ , the average of all individuals if they could be measured. The average (expected)

squared deviation between an observation and μ is the population VARIANCE. A related measure is the squared deviation between observations made and the sample MEAN, which is represented by the SAMPLE variance. That is, deviations are defined in terms of the sample mean rather than population mean.

In a more general context, deviation refers to the discrepancy between a sample of observations and some model that supposedly represents the process by which all data are generated. Perhaps the simplest and best-known example is determining whether observations were sampled from some specified DISTRIBUTION, with a NORMAL DISTRIBUTION typically being the distribution of interest. The most commonly used measure of deviation between the distribution of the observed values, versus the hypothesized distribution, is called a Kolmogorov distance (e.g., see Conover, 1980). The method has been extended to measuring the overall deviation between two distributions (Doksum & Sievers, 1976), which can result in an interesting perspective on how two independent groups compare. (For details and appropriate software, see Wilcox, 2003.)

The notion of deviation arises in several other situations. In regression, for example, a common model is $Y = \beta_0 + \beta_1 X + \varepsilon$, where ε is usually taken to be a variable having a mean of zero. So, the model is that if we are told that $X = 10$, for example, the mean of Y is equal to $\beta_0 + \beta_1 X = \beta_0 + \beta_1 10$, where, typically, β_0 and β_1 are unknown parameters that are estimated based on the sample of observations available to us. In essence, the model assumes a linear association between the mean of Y and X . Here, deviation might refer to any discrepancy between the assumed model and the data collected in a particular study. So, for example, if Y represents cognitive functioning of children and X represents level of aggression in the home, a common strategy is to assume that there is a linear association between these two measures. A fundamental issue is measuring the deviation between this assumed linear association and the true association in an attempt to assess the adequacy of the model used to characterize the data. That is, is it reasonable to assume that for some choice of β_0 and β_1 , the mean of Y is equal to $\beta_0 + \beta_1 X$? Measures of how much the data deviate from this model have been devised, one of which is essentially a Kolmogorov distance, and they can be used to provide empirical checks on the assumption of a linear model (e.g., Wilcox, 2003).

Yet another context where deviation arises is categorical data. As a simple example, consider a study

where adult couples are asked to rate the effectiveness of some political leader on a 5-point scale. A possible issue is whether a woman's response is independent of her partner's response. Now, the strategy is to measure the deviation from the pattern of responses that is expected under independence. More generally, deviation plays a role in assessing goodness of fit when examining the plausibility of some proposed model (e.g., see Goodman, 1978).

—Rand R. Wilcox

REFERENCES

- Conover, W. J. (1980). *Practical nonparametric statistics*. New York: Wiley.
- Doksum, K. A., & Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63, 421–434.
- Goodman, L. A. (1978). *Analyzing qualitative/categorical data*. Cambridge, MA: Abt.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.

DIACHRONIC

Diachronic features describe phenomena that change over time, implying the researcher employs a chronological, historical, or longitudinal perspective (cf. SYNCHRONIC).

—Tim Futing Liao

DIARY

Diaries are a natural method of data collection for obtaining personal data on a regular (usually day-to-day) time basis (Corti, 1993). They are well-adapted to getting contextualized, detailed, sequential information, and, compared to QUESTIONNAIRES, data from diaries are usually more reliable because they minimize problems of bias in retrospective recall and (being self-administered) can ensure ANONYMITY and freedom from INTERVIEWER EFFECTS. However, there can be problems associated with selection or volunteer bias. Consequently, incentives are often paid to motivate diarists to continue for the full term of their diary-keeping. Diaries are also used in conjunction with

follow-up interviews to expand on episodic behavior (Zimmerman & Wieder, 1977).

FORMS OF DIARY

The diary method has a range of structured forms:

- *Unstructured diaries* have little or no format specified. They are often unsolicited productions, and the diarist decides on what is significant to report. Free-format diaries are autobiographical, in natural language (or in code, like Pepys' diary), and more a qualitative resource than a mode of systematic data collection.
- *Semistructured diaries* have a specific form or schema imposed by the researcher, and they are usually solicited. This format allows diaries to be directly compared, but also allows the diarists to state the events in their own words and expand on them. The resulting data often integrate qualitative and quantitative modes.
- *Structured diaries* are more akin to the life history in the sense that the units of time are fixed, and two crucial components of the data are (a) what the specific activity is (relevant activities are usually prespecified and precoded), and (b) how long the given event or activity lasted. This generates directly comparable quantitative data. The most common example is the time-use diary (Juster & Stafford, 1985).

The length of time covered by diaries varies considerably. Time-use studies tend to last only a few days, but other forms are usually timed to cover significant events or provide stable estimates of behavior and tend to be between 1 week and 1 month in length. The diary method has been revolutionized in recent years by the Internet. The online diary has become very popular and covers the range of diary forms outlined above.

APPLICATIONS OF DIARY METHODS

Household surveys often include a diary component, and market researchers and economists use diaries for estimating patterns of consumer expenditure. They have also been used extensively in health and illness studies (Verbrugge, 1980), in diet and nutrition, in researching alcohol and recreational drug use, and in compliance with drug regimes. Sexual behavior has been studied by individual diary-keeping methods at least since the original Kinsey studies in the 1930s and,

more recently, in studies of risk behavior among men who have sex with men (Coxon, 1996).

—Anthony P. M. Coxon

REFERENCES

- Corti, L. (1993). Using diaries in social research (*Social Research Update*, Iss. 2). Guildford, UK: University of Surrey, Department of Sociology.
- Coxon, A. P. M. (1996). *Between the sheets: Sexual diaries and gay men's sex in the era of AIDS*. London: Cassell.
- Juster, F. T., & Stafford, F. P. (1985). *Time, goods and well-being*. Ann Arbor, MI: Institute for Social Research.
- Verbrugge, L. (1980). Health diaries. *Medical Care*, 18, 73–95.
- Zimmerman, D. H., & Wieder, D. (1977). The diary-interview method. *Urban Life*, 5, 479–498.

DICHOTOMOUS VARIABLES

A dichotomous VARIABLE is one that takes on one of only two possible values when observed or measured. The value is most often a representation for a measured variable (e.g., age: under 65/65 and over) or an attribute (e.g., gender: male/female).

If the dichotomous variable represents another variable, that underlying variable may be either observed or unobserved. For example, a dichotomous variable may be used to indicate whether a piece of legislation passed. The dichotomous variable (pass/fail) is a representation of the actual, and observable, vote on the legislation. In contrast, each legislator's vote is the result of an individual, and unobserved, PROBABILITY distribution between voting either for or against the legislation.

Dichotomous variables are most commonly measured using 1 and 0 as the two possible values. The use of 1 and 0 usually has no specific meaning relating to the variable itself. One could just as easily choose 2 and 0, or 1 and –1 as the two values. Even so, the 1, 0 coding does have the advantage of showing, when cases are summed and averaged, the proportion of cases with a score of 1. The choice and assignment of 1 and 0 is usually based on ease of interpretation and the nature of the specific research question.

Dichotomous variables may be used as either DEPENDENT (ENDOGENOUS) or INDEPENDENT (EXOGENOUS) VARIABLES. Many research questions in the social sciences use dichotomous dependent variables. Such questions often involve individual binary choices

or outcomes (e.g., to purchase or not, to vote for or against, to affirm or reverse). Estimating models with dichotomous dependent variables often requires the use of nonlinear techniques because linear models (such as ORDINARY LEAST SQUARES) may be unrealistic on a theoretical level and produce inefficient or inconsistent results.

When used as independent variables, dichotomous variables are often referred to as “controlling” or “DUMMY” VARIABLES. Unfortunately, referring to dichotomous independent variables in this way may lead to inadequately justifying the measurement of the variable or its inclusion in the model. As with any other variable, a dichotomous independent variable should have a sound theoretical basis and be properly operationalized.

As a cautionary note, one must not be overly eager to “dichotomize” variables (i.e., convert a measurable underlying variable to dichotomous form) for three reasons. First, such conversions may not be theoretically justifiable. The popularity of estimation techniques such as PROBIT ANALYSIS and LOGIT caused some to unnecessarily convert continuous dependent variables to dichotomous form. Second, even when theoretically justified, such conversions necessarily result in a loss of information. It is one thing, for example, to know that the Supreme Court affirmed a decision, but quite another to know whether it did so by a 9–0 or 5–4 vote. Third, forcing a variable into just two categories may not always be easy or appropriate if there is a nontrivial residual category. For example, if one measures handedness dichotomously (left/right), the number of ambidextrous individuals or those with arm injuries that required converting to their other hand may pose coding or analytical problems.

—Timothy M. Hagle

REFERENCES

- Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage.
- Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.

DIFFERENCE OF MEANS

Often, social scientists are interested in finding out whether the means of two social groups are identical,

for example, whether the income levels are the same between two social classes or between women and men. This requires SIGNIFICANCE TESTING, and the NULL HYPOTHESIS is that there exists no difference between the two means μ_1 and μ_2 . When only two group means are involved, a simple *T*-TEST is the most often applied procedure, which is also applied to testing DIFFERENCE OF PROPORTIONS, whereas ANALYSIS OF VARIANCE is a natural choice when multiple groups are involved to detect BETWEEN-GROUP DIFFERENCES. A closely related procedure is MULTIPLE CLASSIFICATION ANALYSIS. There are also various MULTIPLE COMPARISON procedures available for dealing with different situations, such as the BONFERRONI TECHNIQUE for related or dependent significance tests. In sum, testing *difference of means* is just one of the many ways of making STATISTICAL COMPARISON.

—Tim Futing Liao

DIFFERENCE OF PROPORTIONS

When people are divided into groups, different proportions of each group will display certain traits, attitudes, or outcomes. For example, different proportions of men and women vote for Democrats, different proportions of Americans and Japanese suffer from breast cancer, and different proportions of mainline and fundamentalist Protestants approve of abortion for unmarried women.

The difference between the abortion attitudes of mainline and fundamentalist Protestants may be estimated using data from the U.S. portion of the 1990 World Values Survey, summarized in the following table:

	Mainline	Fundamentalist	Totals
Approve	172 (n_{11})	23 (n_{12})	195 ($n_{1.}$)
Disapprove	404 (n_{21})	125 (n_{22})	529 ($n_{2.}$)
Totals	576 ($n_{.1}$)	148 ($n_{.2}$)	724 ($n_{..}$)

NOTE: To simplify the example, we treat the World Values Survey like SIMPLE RANDOM SAMPLING. The actual survey design was more complicated.

Of $n_{.1} = 576$ mainline Protestants, $n_{11} = 172$ approved of abortion for unmarried women, a proportion of $p_1 = 172/576 = .2986$. Of $n_{.2} = 148$

fundamentalist Protestants, $n_{12} = 23$ approved of abortion for unmarried women, a proportion of $p_2 = 23/148 = .1554$. Among the surveyed Protestants, then, the *difference of proportions* was $p_1 - p_2 = .1432$. The surveyed mainline Protestants were 14.32% more likely to approve of abortion for unmarried women.

If we wish to generalize from this sample to the U.S. Protestant population, we need to know the SAMPLING DISTRIBUTION for the difference of proportions. In large samples, the difference of proportions $p_1 - p_2$ approximates a NORMAL DISTRIBUTION, whose STANDARD DEVIATION is known as the STANDARD ERROR for the *difference of proportions*.

Different standard errors are used for CONFIDENCE INTERVALS and HYPOTHESIS tests. For confidence intervals, the most common standard error formula is

$$s_{p_1-p_2} = \sqrt{\frac{p_1(1-p_1)}{n_{.1}} + \frac{p_2(1-p_2)}{n_{.2}}},$$

where

$$p_1 = \frac{n_{11}}{n_{.1}} \quad \text{and} \quad p_2 = \frac{n_{12}}{n_{.2}},$$

and the corresponding $(1 - \alpha) \times 100\%$ confidence interval is

$$p_1 - p_2 \pm z_{1-\alpha/2} s_{p_1-p_2},$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2) \times 100$ th PERCENTILE of the standard NORMAL DISTRIBUTION.

Although popular, this confidence interval has poor coverage, especially when the table counts n_{ij} are low. Better coverage, especially for low counts, is obtained by adding 1 to each n_{ij} and 2 to each $n_{.j}$ (Agresti & Cato, 2000). The improved standard error is

$$\tilde{s}_{p_1-p_2} = \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_{.1}+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_{.2}+2}},$$

where

$$\tilde{p}_1 = \frac{n_{11} + 1}{n_{.1} + 2} \quad \text{and} \quad \tilde{p}_2 = \frac{n_{12} + 1}{n_{.2} + 2}$$

and the improved $(1 - \alpha) \times 100\%$ confidence interval is

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{1-\alpha/2} \tilde{s}_{p_1-p_2}.$$

For the Protestant data, $\tilde{p}_1 - \tilde{p}_2 = 0.1393$, $\tilde{s}_{p_1-p_2} = 0.0355$, and $z_{.975} = 1.96$, so a 95% confidence

interval runs from 0.0698 to 0.2088. In other words, the population of mainline Protestants is probably between 6.98% and 20.88% more likely to approve of abortion for unmarried women. (Sometimes, the confidence limits calculated from this formula exceed the lower bound of $-p_1$ or the upper bound of $1 - p_1$. On such occasions, it is customary to adjust the confidence limits inward.)

If we wish to test the NULL HYPOTHESIS of equal proportions, we use a slightly different standard error. Under the null hypothesis, we regard the surveyed Protestants as a single sample of $n_{..} = 724$ people, of whom $n_{1.} = 195$ approve of abortion for unmarried women, a proportion of $p = 195/724 = 0.2693$. In this setting, an appropriate standard error formula is

$$\hat{s}_{p_1-p_2} = \sqrt{p(1-p) \left(\frac{1}{n_{1.}} + \frac{1}{n_{2.}} \right)},$$

where

$$p = \frac{n_{1.}}{n_{..}}.$$

For the Protestant data, $\hat{s}_{p_1-p_2} = 0.0409$. An appropriate test statistic is

$$z = \frac{p_1 - p_2}{\hat{s}_{p_1-p_2}},$$

which, under the null hypothesis, follows an approximate standard normal distribution. For the Protestant data, $z = 0.1432/0.0409 = 3.503$, which has a two-tailed P VALUE of 0.00046. So we reject the null hypothesis that mainline and fundamentalist Protestants are equally likely to approve of abortion for unmarried women.

When squared, this z statistic becomes the CHI-SQUARE (χ^2) TEST statistic used to test for association in a 2×2 contingency table. For the Protestant data $\chi^2 = z^2 = 12.27$, which, like the z statistic, has a p value of 0.00046.

The z and χ^2 tests are considered reasonable approximations when all table counts n_{ij} are at least 5 or so. For tables with lower counts, z and χ^2 should be avoided in favor of Fisher's exact test (Hollander & Wolfe, 1999).

—Paul T. von Hippel

REFERENCES

- Agresti, A., & Cato, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician*, 54(4), 280–288.
- Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences*. Upper Saddle River, NJ: Prentice Hall.
- Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd ed.). New York: Wiley.

DIFFERENCE SCORES

A difference score is the difference of two values, usually two values on the same variable measured at different points in time. For example, suppose variable X is an attitude scale on Aggressiveness and is administered to subjects at Time 1, and then again 6 months later at Time 2. Then, $X_1 - X_2$ equals the difference score, or change, in Aggressiveness from one time to the next. Some variables measured repeatedly over time, such as number of homicides in a nation, tend to exhibit an upward drift. Such variables may be converted to difference scores in order to flatten out that trend.

—Michael S. Lewis-Beck

See also CHANGE SCORES

DIMENSION

A term used by Paul Lazarsfeld (1958) to refer to the different aspects that a CONCEPT may assume. Many concepts are regarded as multidimensional in the sense that it may be possible to specify on a priori grounds different components of those concepts. *Dimension* also refers to the directional spread in a data MATRIX or a CONTINGENCY TABLE, such as a row (vector or variable), a column (vector or variable), etc.

—Alan Bryman

REFERENCE

- Lazarsfeld, P. (1958). Evidence and inference in social research. *Daedalus*, 87, 99–130.

DISAGGREGATION

Disaggregation is the breaking down of aggregates into smaller and smaller units. Data on geographic units, say, the American states, may be disaggregated to congressional districts, or to the still-smaller unit of the county. Aggregated data are often aggregations from a smaller to a larger geographic unit, as in the foregoing example. However, it may refer to human populations. For example, behavior observed at the group level, such as cooperation in a work organization, might be disaggregated to the individual level, such as the cooperativeness of the individual worker. Or, as another example, U.S. presidential voting behavior might be studied at the disaggregated level of the individual voter, or at the highly aggregated level of the presidential election outcome in the nation as a whole.

—Michael S. Lewis-Beck

DISCOURSE ANALYSIS

There is a wide range of theory and method that uses the term *discourse analysis* (henceforth DA), largely because the varieties of DA derive from different academic disciplines. All of them are concerned with the structures and functions of discourse, or talk and text. However, most varieties of DA have little obvious connection with each other. We focus here on the kind of DA that is closely tied to the analysis of social actions and interaction, and that has developed primarily within the social sciences, and social psychology in particular. This excludes the cognitive psychology of “discourse processing,” as well as developments within linguistics that extend grammatical and pragmatic analysis beyond the boundaries of single sentences. An impression of the broad range of kinds of DA can be gleaned from Teun Van Dijk’s (1985) four-volume collection, as well as from other entries in this encyclopedia, particularly those on CRITICAL DISCOURSE ANALYSIS, which is grounded in linguistics; CONVERSATION ANALYSIS, which is based in sociology; and NARRATIVE ANALYSIS, which is derived from linguistics and literary criticism.

In their influential work *Discourse and Social Psychology*, Jonathan Potter and Margaret Wetherell

(1987) defined a theoretical and methodological approach to discourse that focuses on everyday descriptions and evaluations, or “versions,” of things. Versions have three major features: *construction*, *function*, and *variability*.

1. *Construction* has two senses. Discourse is both constructed (built from ready-made linguistic resources) and constructive (offering particular versions of the world, as distinct from alternative versions).

2. Discourse is always performative, or *functional*. That is, things said or written in everyday talk and text are invariably produced with regard to some context of interaction or argument, where they perform actions such as evaluating, criticizing, requesting, confessing, claiming, defending, refusing, and so on.

3. Versions are *variable*, which is to say that we should not expect people to be consistent. It is this latter property, variability, that generated a profound critique of traditional theory and methods concerning the nature and measurement of attitudes, and also the conduct and interpretation of interviews.

The three features are mutually implicated. Discourse is *constructive* of whatever it describes, in that an indefinitely extensive range of alternative descriptions is always possible, without their becoming simply or objectively false. This permits the choice of any particular description to be *functional*, or action-performative, not only in clear cases, such as requests and questions, but also in the case of ostensibly simple, straightforward, factual descriptions. Indeed, DA’s major analytic concern has been with the constructive and functional work done by factual descriptions. Furthermore, given the functional work that versions perform, *variability* across versions, even within the talk of the same speaker, can be expected on the basis that different versions may be doing different things on different occasions. Again, this principle has generated far-reaching critiques of other methods, including traditional uses of INTERVIEWS and QUESTIONNAIRES, in which variability is systematically avoided or removed in favor of defining a person’s consistent attitude or understanding.

HISTORICAL DEVELOPMENT

DA’s sources include linguistic philosophy, POSTSTRUCTURALISM, RHETORIC, ETHNOMETHODOLOGY,

and conversation analysis. These coalesced in an approach to discourse that focuses on its place in social practices and its role in defining, legitimating, and undermining factual versions of the world. Its immediate origin was in the sociology of science. Nigel Gilbert and Michael Mulkay (1984) had a problem finding consistency across the things scientists said in interviews, in their informal talk, and in the accounts given in scientific papers of the grounds on which knowledge claims are made and refuted. Two crucial analytic moves were necessary. First, rather than using current scientific consensus as their own criterion of truth and error, they decided to treat all notions of truth and error as participants' constructions, and to focus on analyzing how scientists produced them. Second, rather than worrying about how to eliminate inconsistency, they chose to make inconsistency, or variability, the central phenomenon. Two contrasting INTERPRETATIVE REPERTOIRES were identified in scientific accounts: an *empiricist repertoire* that accounted for scientific truth, and a *contingent repertoire* that called upon personal and social factors and foibles in accounting for error. The study emphasized how both kinds of accounts, rather than just the standard empiricist repertoire, were essential elements in scientific explanation.

Potter and Wetherell's contribution was to extend these developments in the sociology of science into a broadly applicable theory and method for analyzing discourse in general. They first extended the principles of construction, function, and variability, and of interpretative repertoires, to the study and critique of attitude theory and measurement in social psychology. Rather than looking for coherent and consistent psychological states such as "attitudes," which might be assumed to be expressed in things people say, Potter and Wetherell, along with other discourse-based critics such as Michael Billig, focused on the ways in which descriptions and evaluations are produced in talk. Again, variability and the functional, action-performative nature of discourse came to the fore, and criticisms were produced of the older established methods by which such things are usually hidden from view in pursuit of consistency, reliability, or measures of central tendency.

Recent work in DA has taken two main directions, reflected in different methodological emphases and research topics. One direction has been to develop the notion of "interpretative repertoire" into a concept capable of linking the details of discourse

practices, generally made available through transcripts of interview talk, to larger historical and ideological themes. Repertoires are considered to provide the resources on which people draw when constructing versions of people and events in the world. This strand of DA (e.g., Wetherell & Potter, 1992) focuses on controversial topics with clear ideological import, such as race, gender, and discriminatory practices. It is influenced by Michel Foucault's concept of discourses as historically evolved concepts and frameworks of explanation related to the growth of institutions such as the law, education, medicine, and psychiatry, tracing their infusion into commonsense language and ideology. In contrast to Foucault's discourses, interpretative repertoires are intended to be more analytically tractable, being closer to discourse *practices*, or the actions and interactional work performed by actual instances of talk and text. The second direction in DA's development is called *discursive psychology*, which, taking its major analytic impetus from conversation analysis, focuses on the close analysis of recordings of everyday, naturally occurring talk and, to a lesser extent, textual materials such as newspaper reports.

DISCURSIVE PSYCHOLOGY

Discursive psychology (henceforth DP) is the application of discourse analytic principles to psychological themes and concepts (Edwards, 1997; Edwards & Potter, 1992). In spite of the focus on psychology, however, its range of relevance and application is very broad. This is because psychological themes are pervasive in the workings of commonsense accountability. In everyday talk, and across a range of institutional settings, there is a prime concern with what people know, think, feel, understand, want, intend, and have in mind, as well as their personal dispositions, and so forth. Indeed, such themes are often defining features of the normative business of institutional settings, such as in education, in courtrooms, and in psychotherapy and counseling, where what people think, understand, and remember are central concerns.

There are three major strands of research in DP:

1. Respecification and critique of psychological topics and explanations, which reworks notions such as "memory" and "emotion" as discourse practices
2. Investigations of how everyday psychological terms (the "psychological thesaurus") are used

3. Studies of how psychological themes are handled and managed in talk and text, without having to be overtly labeled as such

The use of the term *psychological* here does not imply a treatment of these themes and topics as rooted in psychological processes and explanations. Rather, DP avoids psychological theorizing in favor of the pragmatics of social actions. All three strands tend to be relevant in actual studies, where the emphasis is on empirical analysis of everyday and institutional talk and text.

ANALYTICAL PRINCIPLES

There is a range of analytical principles that guides empirical work in DP. The following is a representative rather than definitive or exclusive list.

1. There is a preference for *naturally occurring* data, or discourse produced as part of everyday and institutional life, rather than data obtained through, say, research interviews. This is not crucial, but it reflects the principle that discourse performs social actions; a lot of social science's "qualitative" data are interview talk produced aside from, and somewhat reflective on, the lives discussed.

2. Ask not what state of mind the talk/text expresses, nor what state of the world it reflects, but what *action* is being done by saying things that way.

3. Look for *participants' concerns*; their categories, concepts, the things they are dealing with. Look for how they use psychological concepts, or orient to psychological concerns. DP is not, at any stage, looking for evidence of underlying psychological processes.

4. For any issue that the analyst might bring to the data, try *topicalizing* it instead. That is, try seeing to what extent it is something that the participants themselves handle or deal with in some way.

5. Focus on *subject-object* relations (or "mind-world" relations). Look for how descriptions of people and their mental states are tied to, or implied by, descriptions of actions, events, and objects in the external world, and vice versa.

6. Examine how the *current speaker/writer* attends reflexively to his or her own subject-object issues: his or her grounds for knowing and telling things; how he or she deals with the possibility of not being believed, or of being considered biased or emotionally involved.

7. For any content of talk, ask *how*, *not why*, it is said. Ask "What does it do, and how does it

do it?" "Why" questions often depend on normative assumptions about mind, language, and social settings, and they are often best translated into "how" questions. So, instead of asking "Why did X say that?" we can ask, "Did X say that in some way that attends to her possible motives or reasons for saying it?"

8. Analyze *rhetorically*. Ask "What is being denied, countered, forestalled, etc., in talking that way?"

9. Analyze *semiotically*. In the case of DA (see also SEMIOTICS), this means asking, "What is *not* being said here that could have been said by using closely similar words or expressions?" The principle is that language is a system of differences, such that all words, all details, have meanings because there are alternatives. The selection of a particular word or expression is crucial, and we can get to it analytically by imagining plausible alternatives and by looking to see what alternative descriptions may actually be in play. This is a method for seeing the specificity of what is said, not a claim that any such alternatives are actually entertained by speakers.

10. Analyze *sequentially*. For any stretch of discourse, look at the immediately prior and subsequent discourse, or turns at talk, to see what the content of the current turn or section is dealing with and making relevant. This is the major principle of conversation analysis, enshrined in the PROOF PROCEDURE, and it is an essential ingredient in DP for all the same reasons. What we are analyzing is not a collection of speakers' thoughts being put into words, like quotations, but a sequence of actions being performed in sequentially relevant ways.

11. When recurrent patterns are found, look for *deviant cases*, which are examples that do not seem to fit the developing analysis, and see if the analysis needs to be changed or the phenomenon redefined. Again, DEVIANT CASE ANALYSIS is a principle derived from conversation analysis.

AN ILLUSTRATION

Discourse analysis deals with bulky and varied data, not coded and reduced to categories, as in CONTENT ANALYSIS, but retaining its full textual, recorded, or transcribed form, as in conversation analysis. One consequence of that is that all of its principles cannot easily be demonstrated with brief examples. Nevertheless, we can illustrate the kind of data and analysis involved.

The following brief extract comes from a relationship counseling session in which husband Jeff and wife Mary are talking with a counselor about their

marital difficulties. Relevantly to the extract, Mary has had an extramarital affair and is talking about Jeff's reactions to being told about it (see Edwards, 1997, for more details).

1. Mary: U::m (1.0) and then::, (.) obviously
2. you went through your a:ngry stage,
3. didn't you?
4. (.)
5. Ve:ry upset obviously, hh an:d uh, (0.6)
6. we: started ar:guing a lot, an:d (0.6)
7. just drifted awa:y.

Mary describes Jeff's reactions as "angry" (line 2) and "very upset" (line 5). Drawing on the *constructive*, *rhetorical*, and *semiotic* principles, note that these categories characterize Jeff's reactions as angry rather than, say, as having arrived at a condemnatory judgment of his wife's actions and character. They define his reaction as emotional, and as a specific kind of emotion. "Angry" sets up various possibilities that a cognitive "judgment," for example, would not imply. Emotion discourse often trades on its commonsense irrationality, for example, which can be used to lead us away from making inferences about Mary and focus more on Jeff's state of mind. Similarly, a view, opinion, or judgment can (normatively) be more enduring than an emotional reaction, such as anger, and less expected to change. These are conceptual and rhetorical possibilities at this point in the analysis. Their significance will depend on what Mary and Jeff go on to say, because it is their significance *for them* that we are trying to find.

The description "your angry stage" starts to exploit a notion of anger as a temporary state that has its proper occasions and durations. It is a description that implies a kind of script (a routine, expectable sequence of events) for emotional reactions. For example, while acknowledging that Jeff's anger is normatively expectable, the idea of an angry *stage* projects the implication that it should not endure for an unreasonably long time. Mary makes rhetorical room here for something that she actually does go on to develop after this extract, which is the notion that Jeff's reactions are starting to become the problem that they have in their relationship. His enduring emotional reaction can start to become the topic for the counseling to focus on—his inappropriate feelings, rather than her unfaithful behavior and possible proclivities.

Therefore, the expression "your angry stage" is constructive, rhetorical, and performative. It is not just descriptive of Jeff (as an object in the world), nor

merely an expression of Mary's mental understanding of him (her thoughts put into words). Rather, it works to *define* the problem they have as one residing in Jeff, while at the same time avoiding overt blame of him by Mary. Indeed, the term "your" angry stage, along with the normative notion "stage" and the expression "obviously" (line 1), present his reaction as one to which he is normatively entitled, at least for a limited period.

Note also how the description fits into a narrative sequence. The next thing that Mary recounts (and, by implication, what not only follows but *follows from* Jeff's reactions; see the entry on narrative analysis), is how "we started arguing a lot, and just drifted away" (lines 6–7). Their problems become joint ones, arguments, and a kind of nonagentive, nonblaming, "just" drifting apart. Jeff's "stage" is already past tense ("you *went* through your angry stage, *didn't* you?"), which again implies that his anger ought now to be over. The general direction of Mary's account, produced through its specific details, is to shift our understanding of their relationship troubles from a focus on her extramarital affair toward her husband's persisting emotional difficulties in dealing with it.

—Derek Edwards

REFERENCES

- Edwards, D. (1997). *Discourse and cognition*. London: Sage.
- Edwards, D., & Potter, J. (1992). *Discursive psychology*. London: Sage.
- Gilbert, G. N., & Mulkay, M. (1984). *Opening Pandora's box: A sociological analysis of scientists' discourse*. Cambridge, UK: Cambridge University Press.
- Potter, J., & Wetherell, M. (1987). *Discourse and social psychology: Beyond attitudes and behaviour*. London: Sage.
- Van Dijk, T. A. (Ed.). (1985). *Handbook of discourse analysis: Vols. 1–4*. London: Academic Press.
- Wetherell, M., & Potter, J. (1992). *Mapping the language of racism: Discourse and the legitimization of exploitation*. Brighton, UK: Harvester Wheatsheaf.

DISCRETE

The term refers to a variable such as a CATEGORICAL variable, which groups units into discrete clusters. It is also often used to refer to instances where the units that make up a variable are discrete. Thus, number of children as a variable can only be in discrete units—you cannot have 3.67 children.

—Alan Bryman

See also ATTRIBUTE, CATEGORICAL, NOMINAL VARIABLE

DISCRIMINANT ANALYSIS

The basic idea of discriminant analysis is clearly illustrated by a study from sociologist Fred Provoost (1979). He conducted a scientific investigation into the discrimination between the rich and poor neighborhoods in the Dutch-speaking part of Belgium and the district of Brussels. Six discriminating variables were distinguished (mean income, educational attainment, occupation, housing situation, an index of social and cultural participation, and the presence of services), and these six variables, taken together, were examined for their capability of discriminating significantly between the two kinds of neighborhoods.

Research involving discrimination between social groups consists of three steps. First, and as reliably as possible, we make an a priori classification of groups. Because this classification is the subject of explanation, it represents the (dichotomous) dependent variable Y with categories such as those of rich and poor neighborhoods. Second, statistical data are gathered for a great number of characteristics, their discriminating capacity is determined by means of statistical analysis, and a weighted sum is calculated. These characteristics are the independent variables

$X_1, X_2, \dots$ (discriminating variables); their weights are $k_1, k_2, \dots$, respectively (discriminant weights); and the weighted sum is $f = k_1 X_1 + k_2 X_2 + \dots$ (discriminant function). Third, if the analysis of the second step appears to be successful (significant discriminating capacities of the weighted sum and of each of the characteristics separately), then we have a good instrument for classifying additional units into one of the groups. This last step, classification, offers the opportunity for a well-founded determination of whether the neighborhoods not included in the analysis are rich or poor.

It follows from the foregoing that the structure of discriminant analysis is the same as that of multiple REGRESSION analysis except for one point: The measurement level of the dependent variable Y is now categorical (here even dichotomous) instead of interval. The causal diagram, which illustrates the research problem, is shown in Figure 1.

The units of analysis are districts or neighborhoods. The discriminating variables, mean income (X_1), educational attainment (X_2), $\dots$ up to the level of services (X_6), are measured at the interval or ratio level. As in multiple regression analysis, they can be considered as causal factors in a multicausal model. The dependent variable Y represents the groups. It can be seen as a categorical variable, here a dichotomy with the categories poor and rich. Anticipating the dummy approach, we can use the codes 0 and 1, respectively. We will no longer speak of the prediction of Y scores,

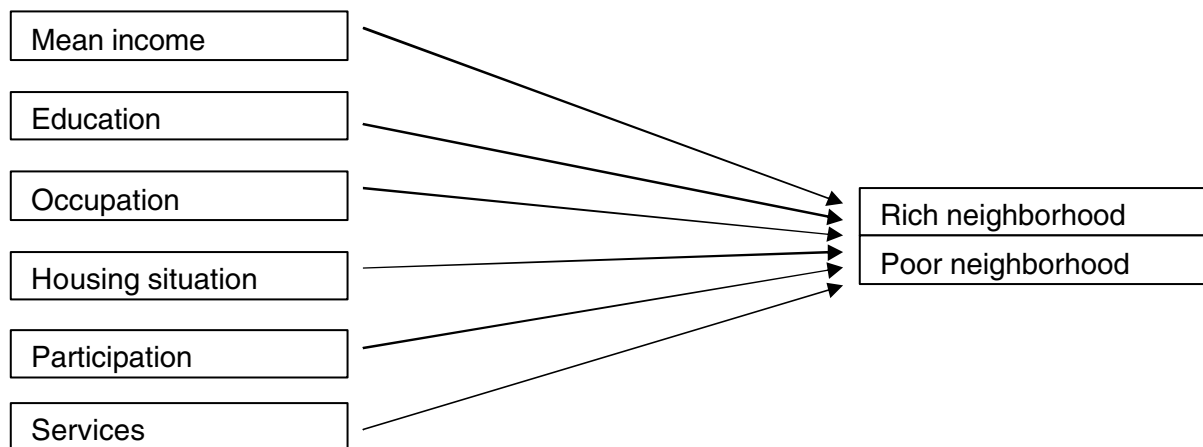


Figure 1

but rather of the classification into one of the two groups. As in multiple regression analysis, the model is also additive. This means that there are no effects of interaction in principle, that is, a weighted sum of X variables is sought, and product terms like X_1X_2 are not considered.

THE MODEL OF DISCRIMINANT ANALYSIS

In discriminant analysis, we want to examine whether a set of variables ($X_1, X_2, \dots$) is capable of distinguishing (i.e., discriminating in the neutral sense) between a number of groups. Therefore, we search for a linear combination of the discriminating variables ($X_1, X_2, \dots$) in such a way that the groups (poor and rich neighborhoods) are maximally distinguished. Such a linear combination is called a discriminant function and generally has the following form:

$$F - \bar{F} = k_1(X_1 - \bar{X}_1) + k_2(X_2 - \bar{X}_2) \\ + \dots + k_p(X_p - \bar{X}_p),$$

or

$$f = k_1x_1 + k_2x_2 + \dots + k_px_p.$$

In this formula, f and x_i are expressed as deviations of mean (small letters). The coefficients k_i are called discriminant weights. The variables x_1 to x_p are the discriminating variables, where p is the total number of X variables.

This can be written much simpler in matrix notation as $\mathbf{f} = \mathbf{X}\mathbf{k}$, in which $\mathbf{f}$ is an $n \times 1$ column vector of discriminant scores, $\mathbf{X}$ is an $n \times p$ matrix of scores of the discriminating variables, and $\mathbf{k}$ is a $p \times 1$ column vector of discriminant weights.

In the case of two groups, there is only one discriminant function. If, on the other hand, several groups were to be compared, then the number of discriminant functions would maximally be equal to $\min(g - 1, p)$. For example, if $g = 4$ groups are analyzed by means of $p = 5$ discriminating variables, then we have $g - 1 = 3$ and $p = 5$, so that the maximum number of discriminant functions is equal to 3.

OBJECTIVES OF THE TECHNIQUE

The purposes of discriminant analysis are threefold. Keeping the example of rich and poor neighborhoods (two groups) in mind and restricting ourselves to two discriminating variables (x_1 and x_2), these purposes can be summarized as follows:

1. We search for a linear function $f = k_1x_1 + k_2x_2$, called a discriminant function, in such a way that the scores of the neighborhoods on this function f (discriminant scores) exhibit the property of maximizing the ratio of between-group and within-group variability. This comes down to the calculation of the discriminant weights k_1 and k_2 and, next, to the calculation of the n discriminant scores.

2. We examine whether the discrimination established by the function f can be generalized to the population. Is there a significant difference between the two group centroids, that is, between the means of the two groups for variables X_1 and X_2 taken together? This comes down to the performance of Hotelling's T^2 test, which is an extension of the simple Student's t -test. In the more general case of three or more groups, this will be Wilks's Λ test.

3. Making use of the discriminant function f , we also want to determine for each neighborhood, and eventually for new neighborhoods, to which of the two groups they belong. Given our knowledge of a neighborhood's score of income and level of services, can we predict whether it is rich or poor? This comes down to determining a cutoff point (in the middle between the two group centroids) and verifying whether the discriminant scores are situated to the left or to the right of this point. If a discriminant score is equal to the cutoff point, then the corresponding neighborhood can be randomly assigned to one of the two groups.

This process of classification of units into groups is mostly considered in the broader context of statistical decision theory and Bayesian statistics.

CALCULATION OF THE DISCRIMINANT FUNCTION f

Function f is a linear combination of x variables: $f = k_1x_1 + k_2x_2 + \dots + k_px_p$. This function f will be calculated in such a way that it shows the highest possible discrimination between the two groups, that is, that the ratio of the between- and within-DISPERSION is maximal. For that purpose, the variation of f is split into a between and within part, and the discriminant weights are calculated in such a way that the ratio between/within is as high as possible.

To solve this problem of maximization, the eigenstructure of a typical matrix will be examined. In discriminant analysis, this typical matrix is $\mathbf{W}^{-1}\mathbf{B}$, which represents the between/within ratio of variations

and covariations. Most of the references at the end of this entry give the details of this eigenstructure examination.

CLASSIFICATION AND PREDICTION

After having found the discriminant function, the expected f value (= discriminant score) of each of the individual neighborhoods can be calculated by substituting the scores of the X variables into the function f . The same operation can also be performed for the group centroids in order to obtain an f score, $\bar{f}_1$, for the centroid of Group 1 and $\bar{f}_2$ for the centroid of Group 2. The point in the middle of these two projections of the group centroids on the f -axis can be chosen as the cutoff point. If the two groups are of equal size, then this cutoff point would lay in the origin, i.e., $f_c = 0$. It will lie to the left of the origin when Group 1 contains fewer neighborhoods than Group 2.

Looking at the discriminant scores of the neighborhoods, we can classify each neighborhood in one of the two groups. Those scoring to the right of f_c ($f_i > f_c$) are assigned to Group 2 and those to the left of f_c ($f_i < f_c$) are classified as Group 1. We can then indicate the predicted group membership. Comparing this with the original Y scores, we can see which percentage of neighborhoods (e.g., 80%) are correctly classified. The discriminant function is not capable of a perfect classification when there is some overlap between the two groups.

SIGNIFICANCE TESTING

We want to examine whether there is a significant difference between the two group centroids, that is, between the means of the two groups for the X variables taken together. This is a test of the global model. In addition, we want to perform separate significance tests, called *univariate tests*, because the X variables are considered one at a time.

For the multivariate test, we make use of Wilks' lambda, Λ , a measure, just like Student's t and Fisher's F , that confronts the ratio of the between-dispersion and the within-dispersion from the sample with this ratio under the null hypothesis. Wilks' Λ is the multivariate analogy of this ratio in its most general form, so that t and F are special cases of Λ . This measure is therefore used for more complex multivariate analysis techniques, such as

MULTIVARIATE ANALYSIS OF VARIANCE AND COVARIANCE, CANONICAL CORRELATION ANALYSIS, and multiple discriminant analysis.

When only two groups are involved, as in our example, then Hotelling's T^2 test can be performed, which is a special case of Wilks' lambda test. T^2 has the nice property that a special matrix is included, $d'C_w^{-1}d$, which is known as Mahalanobis' distance, a generalized distance in which the unequal dispersions of the discriminating variables and their mutual associations are taken into account.

EXTENSIONS

Discriminant weights are commonly computed from the eigenstructure of the data. Other formulations are possible, for example, starting from the total-sample covariance or correlation matrix (C_T or R_T) or from a dummy regression approach obtained by regressing the dummy variable Y on the discriminating variables and leading to the same discriminant weights up to a proportionality transformation. In short, quoting Paul Green, there is an "embarrassment of riches" insofar as computational alternatives for discriminant weights are concerned.

In the foregoing discussion, it was taken for granted that the variables were all measured at the interval or ratio level. It was also assumed that the relationships were all linear. These two requirements can, however, be relaxed. In addition to SPSS, the computer program CRIMINALS, from Leiden University in The Netherlands, performs discriminant analysis, which is nonmetric as well as nonlinear. The interested reader is referred to Gifi (1980).

Linear discriminant analysis is also a technique for which a lot of additional requirements have to be fulfilled; next to linearity, also normality and homogeneity of covariance matrices. In cases in which the covariance matrices do differ significantly, a quadratic discriminant function might be more appropriate. Literature on this topic can be found in Lachenbruch (1975). In case of serious violations of requirements, LOGISTIC REGRESSION is always a good alternative for discriminant analysis.

—Jacques Tacq

REFERENCES

- Cooley, W., & Lohnes, P. (1971). *Multivariate data analysis*. New York: Wiley.

- Gifi, A. (1980). *Nonlinear multivariate analysis*. Leiden, The Netherlands: Leiden University Press.
- Green, P. (1978). *Analyzing multivariate data*. Hinsdale, IL: Dryden.
- Lachenbruch, P. A. (1975). *Discriminant analysis*. New York: Hafner.
- Tacq, J. J. A. (1997). *Multivariate analysis techniques in social science research: From problem to analysis*. London: Sage.
- Van de Geer, J. (1971). *Introduction to multivariate analysis for the social sciences*. San Francisco: Freeman.

DISCRIMINANT VALIDITY

Discriminant (also referred to as divergent) validity is evidence that a measure is not unduly related to other similar, yet distinct, constructs (Messick, 1989). CORRELATION coefficients between measures of a construct and measures of conceptually different constructs are usually given as evidence of discriminant validity. If the correlation coefficients are high, this shows a lack of discriminant validity or weak discriminant validity, depending on the theoretical relationship and the magnitude of the coefficient. On the other hand, if the correlations are low to moderate, this demonstrates that the measure has discriminant validity. Discriminant validity is especially relevant to the validation of tests in which irrelevant variables may affect scores; for example, it would be important to show that a measure of passionate love is unrelated to social desirability.

Other methods of demonstrating discriminant validity are multitrait-multimethod (MTMM) (see MIXED METHODS RESEARCH) and FACTOR ANALYSIS. The MTMM is useful for examining both convergent and discriminant validity. This procedure involves measuring more than one construct by more than one method. For example, a new measure of job satisfaction could be correlated with existing, validated measures of life satisfaction and relationship satisfaction. Self-report data could be obtained, as well as supervisor and partner ratings on each of the three constructs (i.e., three different methods). Theoretically, job, life, and relationship satisfaction are conceptually different constructs; thus, the correlations between these measures would need to be low in order to indicate discriminant validity. Similarly, the correlations between ratings from different observers about the different constructs would be low if discriminant validity is present.

Factor analysis can be used to show that similar but conceptually different constructs are distinct; for example, Pierce, Gardner, Cummings, and Dunham (1989) assessed the discriminant validity of their newly developed organization-based self-esteem measure by factor analyzing it with several affective measures (e.g., organizational commitment). Their findings showed that organization-based self-esteem could be distinguished from affective measures because it loaded on a different factor compared to the other measures. Conversely, factor analysis can also be used to demonstrate lack of discriminant validity. For example, Carless (1998) used confirmatory factor analysis to show that the subscales of a measure of transformational leadership were highly correlated and formed a higher-order construct of transformational leadership. Factor analysis is an essential technique when studying variables that have conceptual overlap and when data are collected from only one source (e.g., an individual respondent). Depression and anxiety, for example, are theoretically distinct constructs, but there is also considerable empirical overlap between them; depressed individuals are also often highly anxious. In order to validate scales, it is important not only to examine the magnitude of the correlation between these variables, but also to demonstrate discriminant validity by factor analysis.

—Sally A. Carless

REFERENCES

- Carless, S. A. (1998). Assessing the discriminant validity of the transformational leadership behaviour as measured by the MLQ. *Journal of Occupational and Organizational Psychology*, 71, 353–358.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.
- Pierce, J. L., Gardner, D. G., Cummings, L. L., & Dunham, R. B. (1989). Organization-based self-esteem: Construct definition, measurement, and validation. *Academy of Management Journal*, 32, 622–648.

DISORDINALITY

Disordinality is an INTERACTION where the RELATIONSHIP between X_1 and Y is different and opposing depending on the value of X_2 . In the usual interaction, the effect of X_1 on Y is merely different,

depending on the value of X_2 . Suppose the question is how Education (X_1 , measured in years) relates to Income (Y , measured in dollars) for men and women (X_2 , scored 1 = male, 0 = female) in a sample from the American labor force. Now regress Y on X_1 , first for the male subsample and then for the female subsample. If the finding is that the REGRESSION slope is significantly steeper for men, we may conclude that there is an interaction, with the impact of education on income dependent on gender. This result would be an ordinal interaction. But suppose instead that our finding was that the regression slope for men was positive, but the slope for women was not only significantly different but negative. That would be a disordinal interaction, where X_1 has a positive effect within one group and a negative effect within the other. It is useful to imagine that, in this instance, the two regression lines, rather than being nonparallel, as in the usual ordinal interaction, actually cross each other.

—Michael S. Lewis-Beck

See also ORDINAL INTERACTION

DISPERSION

Dispersion roughly refers to the degree of scatter or variability in a collection of observations. For example, individuals might differ about whether a political leader is doing a good job; children might respond differently to a method aimed at enhancing reading; and even in the physical sciences, measurements might differ from one occasion to the next because of the imprecision of the instruments used. In a very real sense, it is dispersion that motivates interest in statistical techniques.

A basic issue is deciding how dispersion should be measured when trying to characterize a POPULATION of individuals or things. That is, if all individuals of interest could be measured, how should the variation among these individuals be characterized? Such measures are population measures of dispersion. A related issue is deciding how to estimate a population measure of dispersion based on a SAMPLE of individuals.

Choosing a measure of dispersion is a complex issue that has seen many advances during the past 30 years. The choice depends in part on the goal of the

investigator, with the optimal choice often changing drastically depending on what an investigator wants to know or do. More than 150 measures of dispersion have been proposed, comparisons of which were made by Lax (1985) based on some fundamental criteria that are relevant to a range of practical problems. Although most of these measures seem to have little practical value, at least five or six play an important and useful role.

Certainly the best known measure of dispersion is the population VARIANCE, which is typically written as σ^2 . It is the average (or expected) value of the squared difference between an observation and the population MEAN, μ . That is, if all individuals could be measured, the average of their responses is called the population mean, μ , and if, for every observation, the squared difference between it and μ were computed, the average of these squared values is σ^2 . In more formal terms, $\sigma^2 = E(X - \mu)^2$, where X is any observation we might make and E stands for expected value. The (positive) square root of σ^2 , σ is called the (population) STANDARD DEVIATION. Based on a sample of n individuals, if we observe the values $X_1, \dots, X_n$, the usual estimate of σ^2 is the sample variance:

$$s^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2,$$

where $\bar{X} = \sum X_i / n$ is the sample mean.

For some purposes, the use of the standard deviation, σ , stems from the fundamental result that the probability of an observation being within some specified distance from the mean, as measured by σ , is completely determined under normality. For example, the probability that an observation is within 1 standard deviation of the mean is 0.68, and the probability of being within 2 standard deviations is 0.954. These properties have led to a commonly used measure of EFFECT SIZE (a measure intended to characterize the extent to which two groups differ) as well as a frequently employed rule for detecting OUTLIERS (unusually large or small values). Shortly after a seminal paper by Tukey (1960), it was realized that even very small departures from normality can alter these properties substantially, resulting in practical problems that commonly occur.

Consider, for example, the two distributions shown in Figure 1. One is a standard normal, meaning it has mean 0 and variance 1. The other is not a normal distribution, but rather something called a mixed

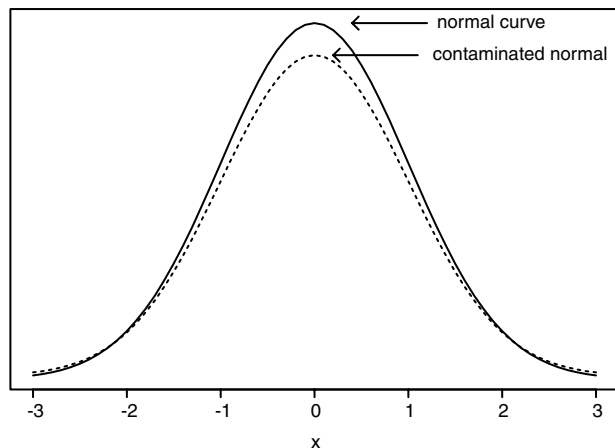


Figure 1

or contaminated normal. That is, two NORMAL DISTRIBUTIONS, both having the same mean but different variances, are mixed together. (A precise description of this particular mixed normal can be found in Wilcox, 2003, but the details are not important here.) The important point is that, despite the obvious similarity between the two distributions, their variances differ substantially: The normal has variance 1 but the mixed normal has variance 10.9. This illustrates the general principle that small changes in the tails of a distribution can substantially affect σ^2 .

Now consider the property that under normality, the probability that an observation is within 1 standard deviation of the means is 0.68. Will this property hold under nonnormality? In some cases, it is approximately true, but in other situations, this is no longer the case, even under small departures from normality (as measured by any of several commonly used metrics by statisticians). For the mixed normal in Figure 1, this probability exceeds 0.999.

The sample variance reflects a similar property: It is very sensitive to OUTLIERS. In some situations, this can be beneficial, but for a general class of problems, it is detrimental. For example, a commonly used rule is to declare the value X an outlier if it is more than 2 (sample) standard deviations from the mean, the idea being that under normality, such a value is unusual from a probabilistic point of view. That is, declare X an outlier if

$$\frac{X - \bar{X}}{s} > 2.$$

This rule suffers from masking, however, meaning that even obvious outliers are not detected because of the sensitivity of s to outliers. Consider, for example, the values 2, 2, 3, 3, 3, 4, 4, 4, 100,000, 100,000. Then, $\bar{X} = 20002.5$, $s = 42162.38$, and $(100,000 - 20,002.5)/s = 1.9$. So, the value 100,000 would not be declared an outlier based on the rule just given, yet it is clearly unusual relative to the other values.

For detecting outliers, two alternative measures of dispersion are typically used. The first is called the median absolute deviation (MAD) statistic. It is the median of the n values

$$|X_1 - M|, \dots, |X_n - M|,$$

where M is the usual median of $X_1, \dots, X_n$. Now X is declared an outlier if

$$\frac{X - M}{\text{MAD}/0.6745} > 2.24.$$

The other alternative measure is the INTERQUARTILE RANGE, which is just the difference between the upper and lower quartiles. (This latter measure of dispersion is used by boxplot rules for detecting outliers.) An important feature of both of these measures of dispersion is that they are insensitive to extreme values; they reflect the variation of the centrally located observations, which is a desirable property when detecting outliers with the goal of avoiding masking.

Another approach when selecting a measure of dispersion is to search for an ESTIMATOR that has a relatively low STANDARD ERROR over a reasonably broad range of situations. In particular, it should compete well with s under normality, but it is desirable to maintain a low standard error under nonnormality as well. So, for example, if it is desired to compare two groups in terms of dispersion, the goal is to maintain high power regardless of whether sampling is from a normal distribution or from the mixed normal in Figure 1. This is roughly the issue of interest in the paper by Lax (1985). Among the many methods he considered, two estimators stand out. The first is called a percentage bend midvariance, and the other is a biweight midvariance. The tedious details regarding how to compute these quantities are not given here, but they can be found in Wilcox (2003) along with easy-to-use software. In terms of achieving a relatively low standard error, the sample standard deviation, s , competes well with these two alternative measures under normality,

but for nonnormal distributions, s can perform rather poorly.

EFFECT SIZE

The variance has played a role in a variety of other situations, stemming in part from properties enjoyed by the normal distribution. One of these is a commonly used measure of effect size for characterizing how two groups differ:

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma_p},$$

where μ_1 and μ_2 are the population means, and where, by assumption, the two groups have equal standard deviations, σ_p . The idea is that if $\Delta = 1$, for example, the difference between the means is 1 standard deviation, which provides perspective on how groups differ under normality. But concerns have been raised about this particular approach, because under nonnormality, it can mask a relatively large difference (e.g., Wilcox, 2003).

STANDARD ERRORS

The notion of variation extends to SAMPLING DISTRIBUTIONS in the following manner. Imagine a study based on n observations resulting in a sample mean, say, $\bar{X}$. Now imagine the study is repeated many times (in theory, infinitely many times), yielding the sample means $\bar{X}_1, \bar{X}_2, \dots$, with each sample mean again based on n observations. The variance of these sample means is called the squared standard error of the sample mean, which is known to be σ^2/n under random sampling. That is, it is the variance of these sample means: $E(\bar{X} - \mu)^2 = \sigma^2/n$. Certainly, the most useful and important role played by the variance is making INFERENCES about the population mean based on the sample mean. The reason is that the standard error of the sample mean, $\sigma/\sqrt{n}$, suggests how inferences about μ should be made assuming normality, but the details go beyond the scope of this entry.

WINSORIZED VARIANCE

Another measure of dispersion that has taken on increasing importance in recent years is the Winsorized variance; it plays a central role when comparing groups based on trimmed means. It is computed as follows. Consider any γ such that $0 \leq \gamma < 0.5$. Let $g = \gamma n$, rounded down to the nearest integer. So if $\gamma n = 9.8$,

say, $g = 9$. Computing a γ -trimmed mean refers to removing the g smallest and g largest observations and averaging those that remain (see TRIMMING). Winsorizing n observations refers to setting the g smallest values to the smallest value not trimmed, and simultaneously setting the g largest values equal to the largest value not trimmed. The sample variance, based on the Winsorized values, is called the Winsorized variance. Both theory and simulation studies indicate that trimming can reduce problems associated with means due to nonnormality, particularly when sample sizes are small. Although not intuitive, theory indicates that the squared standard error of a trimmed mean is related to the Winsorized variance (e.g., Staudte & Sheather, 1990), and so the Winsorized variance has played a role when testing hypotheses based on trimmed means. Also, it has been found to be useful when searching for robust analogs of PEARSON'S CORRELATION (e.g., Wilcox, 2003).

Numerous multivariate measures of dispersion have been proposed as well, several of which currently have considerable practical value. Some of these measures are described in Rousseeuw and Leroy (1987) and Wilcox (2003).

—Rand R. Wilcox

REFERENCES

- Lax, D. A. (1985). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, 80, 736–741.
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression & outlier detection*. New York: Wiley.
- Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.
- Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.

DISSIMILARITY

Dissimilarity is a generic term used to describe measures of similarity or dissimilarity (also called “proximity” measures, especially when used in conjunction with distance models of MULTIDIMENSIONAL SCALING and CLUSTER ANALYSIS). Such measures may

be (a) direct judgments or ratings of similarity made by subjects, or (b) aggregate measures derived from an array of variables. Direct judgments are made between a pair of objects, using a rating scale graded between “totally similar” and “totally dissimilar.” The judgments can then be treated as “quasi-distances” and directly scaled. Aggregate measures differ by LEVEL OF MEASUREMENT and are usually noncorrelational, because for distance model representation, a dissimilarity measure should be nonnegative and symmetric, and it should obey the triangle inequality.

—Anthony P. M. Coxon

REFERENCE

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 5, 5–48.

DISTRIBUTION

Distributions are ubiquitous in social science. They appear in the frameworks for analyzing particular topical domains, where they provide a diversity of shapes to represent the range of possible variability in fundamental quantities; in theories, where they provide basic tools for deriving predictions; and in empirical work, where underlying distributional forms are estimated and distributions serve to characterize operation of the unobservables.

CLASSIFICATION OF DISTRIBUTIONS

Distributions may be classified along several dimensions—continuous versus discrete, univariate versus multivariate, modeling versus sampling—or according to mathematical characteristics of their associated functions. For introduction and comprehensive exposition, see the volumes in Johnson and Kotz’s series, *Distributions in Statistics*, and their revisions (e.g., Johnson, Kotz, & Balakrishnan, 1994), as well as the little handbook by Evans, Hastings, and Peacock (2000).

ASSOCIATED FUNCTIONS

Mathematically specified distributions have associated with them a variety of functions. The most

basic is the distribution function (also known as the cumulative distribution function), which is defined as the probability α that the variate X assumes a value less than or equal to x and is usually denoted $F_x(x)$, or simply $F(x)$. Probably the best known of the associated functions is the PROBABILITY DENSITY FUNCTION, denoted $f(x)$, which, in continuous distributions, is the first derivative of the distribution function with respect to x (and which, in discrete distributions, is sometimes called the probability mass function). For example, in the normal family, the BELL-SHAPED CURVE depicting the probability density function is a more common representation than the ogive depicting the distribution function. Two of the most useful for social science are the quantile function, which, among other things, provides the foundation for whole-distribution measures of inequality, such as Pen’s Parade, and the hazard function.

Of course, all the associated functions are related to each other in specified ways. For example, as already noted, among continuous distributions, the probability density function is the first derivative of the distribution function with respect to x . The quantile function, variously denoted $G(\alpha)$ or $Q(\alpha)$ or $F^{-1}(\alpha)$, is the inverse of the distribution function, providing a mapping from the probability α to the QUANTILE x . Important relations between the two include the following (Eubank, 1988; Evans, Hastings, & Peacock, 2000):

$$\begin{aligned} F[Q(\alpha)] &= \alpha, \text{ for } F \text{ continuous} \\ Q[F(x)] &= x, \text{ for } F \text{ continuous} \\ &\text{and strictly increasing.} \end{aligned}$$

DISTRIBUTIONAL PARAMETERS

Distributions also have associated with them a number of parameters. Of these, three are regarded as basic—the location, scale, and shape parameters. The location parameter is a point in the variate’s domain, and the scale and shape parameters govern the scale of measurement and the shape, respectively. It is sometimes convenient to adopt a uniform notation across distributional families, such as that in Evans, Hastings, and Peacock (2000), which uses lower-case, italicized English letters a , b , and c to denote the location, scale, and shape parameters, respectively. Variates differ in the number and kind of basic parameters. For example, the normal distribution has

two parameters, a location parameter (the mean) and a scale parameter (the standard deviation).

SUBDISTRIBUTION STRUCTURE

Distributions are useful for representing many phenomena of special interest in social science. These include, besides inequality, the subgroup structure of a population. In general, two kinds of subgroup structure are of interest, and they may be represented by versions of the distributional operations of CENSORING AND TRUNCATION. In the spirit of Kotz, Johnson, and Read (1982, p. 396) and Gibbons (1988, p. 355), let censoring refer to selection of units by their ranks or percentage (or probability) points; and let truncation refer to selection of units by values of the variate. Thus, the truncation point is the value x separating the subdistributions; the censoring point is the percentage point α separating the subdistributions. For example, the subgroups with incomes less than \$20,000 or greater than \$80,000 each form a truncated subdistribution; the top 5% and the bottom 10% of the population each form a censored subdistribution.

There is a special link between these two kinds of subgroup structure and Blau's (1974) pioneering observation that much of human behavior can be traced to the differential operation of quantitative and qualitative characteristics. Quantitative characteristics—both cardinal characteristics, such as wealth, and ordinal characteristics, such as intelligence or beauty—generate both truncated and censored subdistribution structures. For example, the subgroups “rich” and “poor” may be generated by reference to an amount of income or by reference to percentages of the population. However, qualitative characteristics—such as race, ethnicity, language, or religion—may be so tightly related to quantitative characteristics that the subgroups corresponding to the categories of the qualitative characteristic are nonoverlapping and thus provide the basis for generating censored subdistribution structures. For example, in caste, slavery, or segmented societies, the subdistribution structure of interest may be a censored structure in which the percentages pertain to the subsets formed by a qualitative characteristic—such as “slave” and “free.”

DISTRIBUTIONS IN THEORETICAL WORK

Often, theoretical analysis of a process or derivation of predictions from a theory makes use of modeling

distributions. This approach makes it possible to investigate the possible variability of a process or, relatedly, the impact of the shape of underlying distributions, such as the income distribution. The protocol involves selecting appropriate modeling distributions and preparing them for use in the particular theoretical problem.

To illustrate, we consider three theoretical problems. The first, based on a theory of choice attributed to St. Anselm, investigates the effect of inequality on own income. The second, based on a theory of status integrating elements pioneered by Berger and his associates (Berger, Fisek, Norman, & Zelditch, 1977), Goode (1978), Sørensen (1979), and Ridgeway (2001), investigates the effect on status of rank on valued quantitative characteristics (Jasso, 2001b). The third, based on a theory of justice and comparison processes, investigates the effects of valuing cardinal versus ordinal goods and of the distributional shapes of cardinal goods and their inequality on the distribution of such outcomes as happiness and the sense of being justly or unjustly rewarded (Jasso, 2001a).

The second and third problems include attentiveness to ranks (always in status theory, sometimes in justice-comparison theory). Ranks, as is well known, are directly represented by the rectangular distribution (the discrete rectangular in problems involving small groups and the continuous rectangular in problems involving large societies).

The first and third problems include distributions of cardinal goods, and these require selection of modeling distributions from among distributions with a positive support. One approach is to consider behavior at the extremes, that is, whether there is a minimum income or a maximum income. This approach leads to four combinations: (a) a variate with nonzero infimum and a supremum, (b) a variate with nonzero infimum and no supremum, (c) a variate with zero infimum and a supremum, and (d) a variate with zero infimum and no supremum. To illustrate, the wealth distribution of a primitive society in good times may be modeled by the first type and in bad times (such as famine) by the third type; the wealth distributions of advanced societies near the end of a war may be modeled by the second type among the winners and by the fourth type among the losers. Table 1 provides example(s) of each of the four combinations of properties. As shown, the Pareto variate, as well as the exponential variate, exemplify the combination of nonzero infimum and no supremum. The lognormal exemplifies the combination of

Table 1 Prototypical Distributions of Material Goods

<i>Infimum</i> > 0	<i>Supremum</i>	
	<i>Yes</i>	<i>No</i>
Yes	Quadratic	Exponential Pareto
No	Power-Function	Lognormal

NOTE: In both the power-function and the quadratic variates, the supremum is also the maximum. The lognormal and the power function have an infimum at zero; the power-function of $c > 1$ also has a minimum of zero.

zero infimum and no supremum. The power-function variate has a supremum but no infimum. Finally, the quadratic, a symmetric variate, has both an infimum and a supremum.

In using distributions to model cardinal quantities in the first and third problems, it is useful to have at hand formulas for the variates' associated functions expressed in terms of two substantively meaningful parameters, the mean and inequality. Accordingly, the first step is to adapt the selected variates so that the versions have two parameters, and these two parameters are the arithmetic mean and a general inequality parameter of which all measures of relative inequality are monotonic functions. Table 2 provides the formulas for the cumulative distribution function, the probability density function, and the quantile function for the five variates highlighted in Table 1.

In the Anselmian problem, the protocol involves setting up the quantile function for each modeling distribution and finding the first partial derivative of own income with respect to inequality. Results include the prediction that the proportion of the population for whom own income increases as income inequality decreases varies greatly—for example, 37% when income is power-function distributed, 50% when income is lognormally distributed, and 63% when income is Pareto distributed. In Anselmian theory, these individuals include the poorest members of the society, and they cannot be characterized as being egoistic or altruistic, for the same choice that increases their own income also decreases income inequality.

In the status and justice-comparison problems, an early task is to find the distributions of status and of the justice-comparison outcomes, respectively. This is done by using the formulas for status and for justice-comparison (provided by the

ASSUMPTIONS of the respective theories) and the formulas for the modeling distributions, and applying change-of-variable, convolution, and other techniques (described in standard sources). In the case of status, it is easy to find the distribution in the case of one-good status because the status formula, already expressed in terms of ranks, is seen by inspection to be that for the quantile function of the positive exponential (Sørensen, 1979). Other status distributions, in cases of two goods simultaneously conferring status, include the Erlang and an as-yet unnamed family (Jasso, 2001b, p. 122). Justice-comparison distributions that have been obtained include the negative exponential, the positive exponential, the normal, the logistic, and symmetric and asymmetric Laplace (Jasso, 2001a, pp. 683–685).

Further analysis of the status and justice-comparison distributions yields a large variety of testable predictions, including, for example, the predictions that the rate of out-migration, the public benefit of religious institutions, and conflict severity among warring subgroups all increase as inequality in the distribution of valued material goods increases.

Of course, an important challenge is to obtain distribution-independent results. For example, use of tools from the study of probability distributions yields the result that two singly sufficient but not necessary conditions for the justice-comparison distribution to be symmetric about zero are that (a) the actual reward and the just reward be identically and independently distributed, and (b) the actual reward and the just reward be identically distributed and perfectly negatively related (Jasso, 2001a, pp. 683–685).

DISTRIBUTIONS IN EMPIRICAL WORK

Social science researchers are constantly concerned about the distributions of their DATA, especially the distribution of the outcome or DEPENDENT VARIABLE. Although there exist various reasons for studying distributions of variables, two major purposes stand out—to gain knowledge through analyzing distributions and to approximate underlying distributions.

1. *Gaining further knowledge.* Often, the basic information of distributional parameters is not sufficient, and researchers resort to graphical means to assist their understanding and possibly further

Table 2 Modeling Distributions for Cardinal Goods and Associated Basic Functions

<i>Variate Family</i>	<i>Cumulative Distribution Function</i>	<i>Probability Density Function</i>	<i>Quantile Function</i>
Exponential $x > \frac{\mu}{c}, c > 1$	$1 - \exp\left[-\frac{cx - \mu}{\mu(c-1)}\right]$	$\frac{c}{\mu(c-1)} \exp\left[-\frac{cx - \mu}{\mu(c-1)}\right]$	$\mu\left[\frac{1}{c} + \frac{c-1}{c} \left(\ln \frac{1}{1-\alpha}\right)\right]$
Lognormal $x > 0, c > 0$	$F_N\left\{\frac{[\ln(\frac{x}{\mu}) + \frac{c^2}{2}]}{c}\right\}$	$\frac{1}{xc\sqrt{2\pi}} \exp\left\{-\frac{(\frac{c^2}{2} + \ln \frac{x}{\mu})^2}{2c^2}\right\}$	$\mu \exp\left[cQ_N(\alpha) - \frac{c^2}{2}\right]$
Pareto $x > \frac{\mu(c-1)}{c}, c > 1$	$1 - \left[\frac{\mu(c-1)}{cx}\right]^c$	$\left[\frac{\mu(c-1)}{c}\right]^c cx^{-c-1}$	$\frac{\mu(c-1)}{c(1-\alpha)^{1/c}}$
Power-Function $0 < x < \frac{\mu(c+1)}{c}, c > 0$	$\left[\frac{xc}{\mu(c+1)}\right]^c$	$\left[\frac{c}{\mu(c+1)}\right]^c cx^{c-1}$	$\frac{\mu(c+1)\alpha^{1/c}}{c}$
Quadratic $\frac{\mu(2+c)}{2}$ $< x < \frac{\mu(2-c)}{2},$ $0 < c < 2$	$\frac{1}{2}\left\{1 - \cos\left[3 \arccos\left(\frac{x-\mu}{\mu c}\right) - 4\pi\right]\right\}$	$\left(-\frac{6}{\mu^3 c^3}\right)x^2 + \left(-\frac{12}{\mu^2 c^3}\right)x$ $+ \frac{3(c^2-4)}{2\mu c^3}$	$\mu\left\{1 + c \sin\left[\frac{\arcsin(2\alpha-1)}{3}\right]\right\}$

NOTE: For all variates, $x > 0$; other restrictions as indicated. The terms $F_N(\cdot)$ and $Q_N(\cdot)$ denote the cumulative distribution function and the quantile function, respectively, of the unit normal variate. Inequality is a decreasing function of c in the Pareto and the power-function variates and an increasing function of c in the exponential, lognormal, and quadratic variates.

operations on the data. These can be a HISTOGRAM before applying a particular CODING scheme, a STEM-AND-LEAF DISPLAY or a BOXPLOT for catching OUTLIERS and potential INFLUENTIAL CASES, a Lorenz curve for illustrating distributional inequality, or a SCATTERPLOT for displaying the distribution of one variable against another. Contemporary statistics offers many useful tools for exploring and examining distributions, especially when STATISTICAL COMPARISON is the aim. For example, the RELATIVE DISTRIBUTION METHOD provides researchers with an invaluable way to compare two income distributions.

2. *Approximating underlying distributions.* It is common practice in the social sciences to apply some form of REGRESSION model to our outcome variables, ranging from a simple LINEAR REGRESSION to a nonlinear random coefficient model. By doing so, we assume we know the underlying mechanism that generates the dependent variable. However, too often, convention governs our modeling attempts, without much due attention to

distributional properties. A little exploration would go a long way in better capturing the underlying distribution. Figure 1 gives the distributions of a sample of women from the 1991–2000 British Household Panel Survey, with a histogram representing the empirical distribution of the data and the lines representing three theoretical distributions. The variable measures depression and records the number of items on which a woman reported “feeling worse” than before.

Clearly, the normal distribution does not represent the data well. The Poisson distribution is better, but among the three candidates considered, the negative binomial distribution outshines the others by a close approximation of the data, hence of the underlying distribution.

FURTHER READING

The volumes in Johnson and Kotz’s series, *Distributions in Statistics*, and their revisions represent an

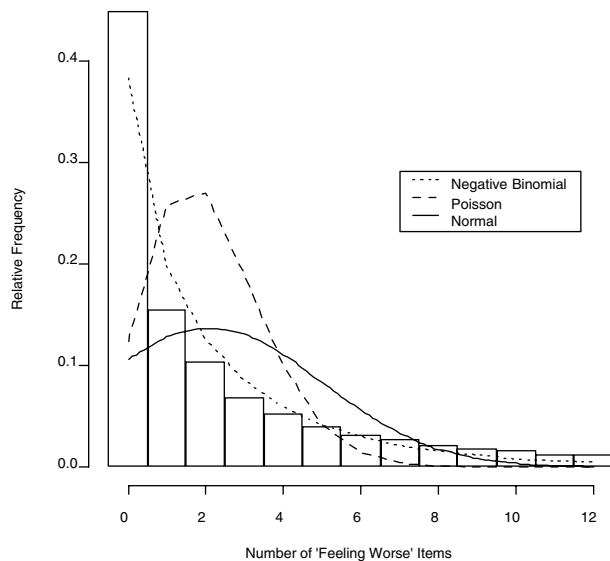


Figure 1 Distributions of GHQ12, 1991–2000 BHPS ($N = 7,855$)

invaluable resource for both theoretical and empirical work in the social sciences.

—Guillermina Jasso and Tim Futing Liao

REFERENCES

- Berger, J., Fisek, H., Norman, R., & Zelditch, M. (1977). *Status characteristics and social interaction: An expectation states approach*. New York: Elsevier.
- Blau, P. M. (1974). Parameters of social structure. *American Sociological Review*, 39, 615–635.
- Eubank, R. L. (1988). Quantiles. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 7, pp. 424–432). New York: Wiley.
- Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical distributions* (3rd ed.). New York: Wiley.
- Gibbons, J. D. (1988). Truncated data. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 9, p. 355). New York: Wiley.
- Goode, W. J. (1978). *The celebration of heroes: Prestige as a control system*. Berkeley: University of California Press.
- Jasso, G. (2001a). Comparison theory. In J. H. Turner (Ed.), *Handbook of sociological theory* (pp. 669–698). New York: Kluwer Academic/Plenum.
- Jasso, G. (2001b). Studying status: An integrated framework. *American Sociological Review*, 66, 96–124.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions, Vol. 1*. New York: Wiley.
- Kotz, S., Johnson, N. L., & Read, C. B. (1982). Censoring. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 1, p. 396). New York: Wiley.
- Ridgeway, C. L. (2001). Inequality, status, and the construction of status beliefs. In J. H. Turner (Ed.), *Handbook of sociological theory* (pp. 323–340). New York: Kluwer Academic/Plenum.
- Sørensen, A. B. (1979). A model and a metric for the analysis of the intragenerational status attainment process. *American Journal of Sociology*, 85, 361–384.

DISTRIBUTION-FREE STATISTICS

In the strictest sense of the phrase, a distribution-free statistic is one whose SAMPLING DISTRIBUTION does not depend on the POPULATION from which a random sample was drawn. That is, the statistic has the same sampling distribution irrespective of the form of the parent population, whether it be the normal, exponential, or any other DISTRIBUTION.

The simplest example of a distribution-free statistic is the SIGN TEST statistic, defined as the number of plus signs among, say, n differences within a set of n paired sample observations that have no ties. The number of plus signs among these differences follows the BINOMIAL DISTRIBUTION with PARAMETER p representing the PROBABILITY of a plus sign and n the sample size, regardless of the form of the distribution from which the random sample of pairs was drawn. Another distribution-free statistic, or at least approximately so, is the standardized mean (or z -SCORE) of a random sample of at least 30 observations drawn from any continuous population with known VARIANCE. According to the CENTRAL LIMIT THEOREM, the sampling distribution of such a standardized sample mean is approximately the normal distribution.

In practice, any distribution-free statistic permits a researcher to carry out a distribution-free test, that is, one that does not require any assumptions about the distribution from which the sample was drawn in order to be regarded as a valid test. A valid test is defined as a test for which the exact probability of a TYPE I ERROR is the stated α value, or one for which the P VALUE is exact. The term *distribution-free test* has now been essentially superseded by the term non-parametric test, which includes not only distribution-free tests but also any statistical test whose NULL

HYPOTHESIS does not include the value of a parameter. This latter type of test includes tests of goodness-of-fit, tests for trend, tests for AUTOCORRELATION, and tests for randomness, among others. Nonparametric statistics include a very large number of methods of statistical analysis that are especially useful for small sample sizes.

—Jean D. Gibbons

DISTURBANCE TERM

Disturbance term refers to the ERROR term in a statistical model. If the disturbance is truly STOCHASTIC, then it is WHITE NOISE.

—Tim Futing Liao

DOCUMENTS OF LIFE

Documents of life are all those documents in which people reveal their social and personal characteristics in ways that are accessible for research. The documents could include diaries; letters; photographs; life stories; inscriptions on tombstones; furnishings; videos; and, these days, personal Web sites. In an early description of the method, humanistic anthropologist Robert Redfield described them as “personal documents,” seeing them as one in which the human and personal characteristics of somebody who is, in some sense, the author of the document find expression, so that through its means the reader of the document comes to know the author and his or her views of events with which the document is concerned (Redfield, in Gottschalk, Kluckhohn, & Angell, 1945, p. vii). They are not the second-order constructs of reality made by social scientists, but first-order accounts that attempt to enter the subjective world of informants, taking them seriously on their own terms.

—Ken Plummer

See also AUTOBIOGRAPHY, AUTOETHNOGRAPHY, DIARIES, LIFE COURSE RESEARCH

REFERENCES

Allport, G. (1942). *The use of personal documents in psychological science*. New York: Social Science Research Council.

Gottschalk, L., Kluckhohn, C., & Angell, R. (1945). *The use of personal documents in history, anthropology, and sociology*. New York: Social Science Research Council.

Hoskins, J. (1998). *Biographical objects: How things tell the stories of people's lives*. London: Routledge.

Plummer, K. (2001). *Documents of life-2: An invitation to a critical humanism*. London: Sage.

DOCUMENTS, TYPES OF

A document is any kind of physically embodied text, and a handwritten or printed text on paper, such as a letter or government report, is the archetypal document. However, the range of documents is much wider than this. We may classify documents as *written* (where the text is visible words), *audio* (where the text is sound that can be accessed through aural channels), and *visual* (where the text is a designed or pictorial representation). A text may be produced with a pen, a pencil, a paint brush, a printing machine, a tape recorder, or a computer, and it may be inscribed on clay, stone, parchment, paper, magnetic disks, or electronic display screens. The full range of documents includes newspapers, diaries, stamps, directories, handbills, maps, photographs, paintings, gramophone records, tapes, and computer files.

Documents are sometimes seen as the particular source material for the historian, but they have a wide range of uses across the social sciences. The classical sociologists made far greater use of documentary sources than they did of survey or observational methods. Karl Marx based *Capital* on his heavy use of the publications of the factory inspectors, Max Weber used religious tracts and pamphlets in *The Protestant Ethic and the Spirit of Capitalism*, and Emile Durkheim drew on a variety of official statistics to produce *Suicide*.

Documents can be classified most usefully in terms of their authorship and the conditions for access to them.

Authorship can be subdivided according to whether documents originate in the personal sphere or the public, official sphere, and official documents can be subdivided according to their origin in state or private bureaucracies. This differentiation rarely can be drawn with any precision, particularly in the pre-modern period. In medieval Europe, for example, a distinction between personal family papers, church papers, and state papers would not make a great deal

of sense because of the highly personalized nature of patrimonial authority.

Access concerns the conditions under which documents may be available to people other than their authors. Documents may be “closed” if they are available only to a very limited group of people, such as those who produce them and their bureaucratic superiors. If documents are “restricted,” they are accessible to a wider group, but under limited conditions determined by their producers or their guardians. The most liberal conditions are “open” access through archiving or publication. Archival access exists where documents have been stored and made available to a wide public: Documents may be available with only minimal bureaucratic restrictions, such as the need to register or to acquire a reader’s ticket (see DATA ARCHIVES, ARCHIVING QUALITATIVE DATA). Where documents are published, access is almost completely open because they are accessible to all who are able to afford them.

These two dimensions can be combined as shown in Table 1 to produce 12 different kinds of documents. Closed personal documents (Type 1) include letters, diaries, household account books, and many other domestic items. These kinds of documents are normally available only to the individual who owns them or to the immediate household that produced them. Personal documents may move from one category of access to another. Some may become available more widely if they are deposited and stored in public records offices (Type 3), and some may be published (Type 4). DIARIES, for example, begin as purely closed documents, but are often produced with the intention of making them available to a wider readership. The records of many landed and wealthy families have been deposited in public archives and so become more easily accessible. Many such documents, however, remain in private hands—if they survive for any period at all—and are closed to public access unless specific permission is granted by the owner (Type 2).

Official documents in the private sphere are those produced by organizations such as businesses, schools, hospitals, and churches. Confidential organizational documents (Type 5) include medical records, records of educational attainment, and company personnel records, all of which are normally available only to those with an administrative or professional responsibility for the people to whom they relate. Restricted organizational documents (Type 6), such as share registers or lists of mortgages, are normally made available to researchers, if at all, when they are no

Table 1 A Typology of Documents

		<i>Authorship</i>		
		<i>Personal</i>	<i>Official</i>	
			<i>Private</i>	<i>State</i>
<i>Access</i>	Closed	1	5	9
	Restricted	2	6	10
	Open-archival	3	7	11
	Open-published	4	8	12

longer of practical business relevance. Specific restrictions may be applied to the consultation and use of the documents. For some, however, there is a legal requirement to deposit them in a public archive (Type 7): Share registers in England and Wales usually must be sent each year to the companies’ registration office. Organizational documents that are published (Type 8) may also be published as a legal or other official requirement, as is the case with the published reports and accounts of companies, but many are published as commercial ventures. Examples include timetables and directories, reference books, newspapers, films (see FILM AND VIDEO IN RESEARCH), and other mass media products.

Government documents, whether local or national, involve the same categories of access and are, perhaps, the single largest type of document available to social researchers. Those with closed access (Type 9) include criminal records and security reports, local authority housing records, and current taxation records. Many are covered by official secrecy legislation that prevents unauthorized disclosure outside the immediate department responsible. Under such legislation, many documents remain closed on a permanent basis, although some papers may be opened up for restricted access (Type 10), and others may become fully available in public archives (Type 11). Those state documents that enter public archives often do so only after a specified period, when they are no longer regarded as confidential. Many state documents are produced for publication (Type 12), and this includes acts of Parliament, reports of royal commissions, statistical reports (see OFFICIAL STATISTICS), and research reports.

The general principles involved in using documents are similar to those in any other area of social research. Nevertheless, the particular features of documentary sources do imply specific judgments and specific techniques. The effective use of documents

depends on them being appraised in terms of four criteria: authenticity, credibility, representativeness, and meaning.

The criterion of *authenticity* requires that documents be assessed for their soundness and authorship. An assessment of soundness means discovering whether the document is an original or a copy and, if a copy, whether it is a copy of an original or a copy of a copy. Copying, whether by handwriting, by re-typesetting, or by photocopying, can result in missing or unreadable text, but even originals may contain typographical errors or be incomplete. An “unsound” document is one that is not close enough to an idealized original because it is corrupted in some way. A researcher may need to reconstruct what a sound document would look like before making use of the unsound one available. The more corrupted a document is, the more difficult such a task will be, and, in extreme cases, it may not be possible to reconstruct a sound version of the original document.

Once soundness has been assessed, authorship can be considered. It is important to know, for example, whether diaries attributed to particular individuals were actually written by the individuals concerned. Issues of forgery or fraudulent authorship can arise, even in apparently straightforward cases. Fraudulent documents may be very important sources of evidence, but the fact that they are fraudulent must be known. Authorship can be assessed through the use of both internal and external evidence. Internal evidence of vocabulary and literary style can be combined with external evidence from chemical tests on paper and ink, and from estimations of plausibility, to arrive at a judgment about the attribution of authorship. In many cases, however, the question of authorship is far from straightforward. Official documents are produced by an administrative machine in which any named person may play a minimal role. Similarly, books and newspapers are the result of a complex division of labor in which the work of named writers is processed by copy editors, subeditors, and editors.

The question of *credibility* concerns the sincerity and accuracy of a document. All documents are selective, because it is impossible to construct accounts that are independent of particular standpoints. Nevertheless, they can be more or less credible as accounts, depending on how sincerely the choice of a point of view is made and whether the account gives an accurate report from that standpoint. The sincerity of authors reflects their motivation and the degree of choice that

they exercise. People write for self-justification, to propagandize or to deceive, for financial gain, or simply to report with as much objectivity as possible, and in each case, it is important to assess the extent to which the document is shaped by these reasons. Official documents, for example, may be presented in the form of “information,” but may actually involve an intent to persuade people of a particular position. Similarly, newspapers are produced by journalists and others who are paid to write marketable material and who may be subject to political pressure from a proprietor.

Even a document that has been sincerely produced may be inaccurate. To assess the accuracy of a report, it is necessary to look at the conditions under which it was compiled and, in particular, how close the author was to the events reported. Accuracy is generally held to be greatest in primary sources—first-hand accounts—because these are felt to minimize any loss of accuracy due to lapses of memory, inadequate records, or ignorance. But even first-hand observers face the problem of recording their observations in a form that they can use at a later time to construct accurate reports. It is generally very difficult, for example, to record conversations verbatim. Any system of recording, especially a mechanical one, is likely to be intrusive and so may be avoided. Observers must often rely on memory, and thus, accuracy of recall is a problem even in primary sources.

Assessing *representativeness* depends upon a knowledge of documents’ survival and their availability. It is important to know whether the documents consulted are representative of all the relevant documents, and this depends upon what proportion of the relevant documents have actually survived and whether they are available for researchers to use. If they are to survive, documents must be stored in some way. This storage can range from deposition in a public archive to dumping in a cardboard box. Not all documents get to be stored. Many public and private documents are destroyed soon after their production, whereas others are stored for a period only to be destroyed at a later date. The increasing use of computers to produce and store documents means that, in many cases, no historical record is produced: Documents are edited and updated in real time, so only the latest revision is ever available. Where computer files are archived in backup form and survive for later research use, it is difficult to know how their contents relate to the original continuous database. Many personal documents are not retained, and those

that survive are likely to do so because their owners have judged them to be out of the ordinary in some way. Only a small proportion of the vast number of documents produced in modern bureaucratic organizations can be retained. Official documents are stored while in current use and may then be weeded out before being transferred to an archive. The number of documents that survive may decrease over time through deterioration and decay or through periodic “clear-outs.” Such losses may sometimes occur at random, but they may also involve unknown systematic processes.

The availability of official documents that have survived is limited by considerations of confidentiality and official secrecy. State documents often enter the public sphere after a particular period of time has lapsed, but some documents may be permanently restricted. Problems may be even greater in the personal and private spheres, and especially with family and household documents, where there are generally no legal requirements about access at all. Personal documents also tend to lack the kind of index or calendar that allows the representativeness of official documents to be assessed. Unless they are stored in family archives—which is unusual except for very wealthy families—private household materials are neither archived nor catalogued.

Although documents must be assessed for their authenticity, credibility, and representativeness, it is not necessary that they actually be fully authentic, credible, or representative. What is important is that the researcher knows that he or she is dealing with documents that are inauthentic, incredible, or unrepresentative to some degree. It is then possible to move on to the question of their meaning.

The *meaning* of documents is an issue that arises at two levels: the literal and the interpretive. A literal understanding of a document involves its physical readability, whether it is in a language that can be read, and such issues as dating. Reading faded or decayed documents may require technologies such as microscopes and infrared light, and the reading of computer files will always require specific hardware and software. The reading of audio and visual materials will routinely require access to particular technologies. Issues of literal meaning are, in principle, resolvable on a practical basis, leaving the researcher with a meaningful text to analyze.

It is at this point that the more fundamental question of interpretive meaning arises, and this is a matter that

raises complex theoretical issues. Interpretation is a hermeneutic task that involves an appreciation of the social and cultural context and forms of discourse that structure a text. It is necessary to grasp the underlying point of view, from which the individual concepts in a text acquire their meaning. The assessment of authenticity, credibility, and representativeness discloses the context from which the document emerged and within which it must be interpreted.

Methods of textual analysis cannot be considered in any detail in this entry, but the two principal approaches are quantitative CONTENT ANALYSIS and qualitative semiotic approaches (see QUALITATIVE CONTENT ANALYSIS). Whereas the former involves the use of methods of counting and statistical analysis, the latter involves methods of decoding and deconstruction. Ultimately, however, all interpretation is a qualitative task, and even a quantitative content analysis rests upon some assessment of significance through an often-implicit semiotic process.

The four criteria of authenticity, credibility, representativeness, and meaning should not be seen as separate and distinct stages in the assessment of documents. They are interdependent, and a researcher cannot adequately consider one without simultaneously considering the others. In this sense, the appraisal of documents is a never-ending process. The interpretive meaning that a researcher aims to produce is a tentative and provisional judgment that must be constantly revised as new discoveries and new problems force a reappraisal of the evidence. An understanding of this is essential if a researcher is to make sensible use of documents.

—John Scott

REFERENCE

- Scott, J. (1990). *A matter of record: Documentary sources in social research*. Cambridge, UK: Polity.

DOUBLE-BLIND PROCEDURE

Blind experimental procedures are designed to eliminate biases produced by expectations of study participants, experimenters, and data analysts. Investigators commonly employ any of three levels of blindness in their experimental designs. In a simple

blind EXPERIMENT, participants are unaware of their assignment to the treatment or to the control group. In a double-blind experiment, both the participant and the experimenter are unaware of the group assignment. Finally, in a triple-blind experiment, the participants, experimenter, and those responsible for data entry and analysis are all unaware of the treatment versus control condition of each study trial.

Blind procedures are necessary because participants respond not only to the experimental stimulus but also to their impression of the experiment's HYPOTHESIS. If the manipulation of the experiment is readily apparent, participants may alter their response in accordance with what they perceive to be the desired response. In this case, the measured effect of the experiment represents an unknown combination of intended treatment effects and the demand characteristics of the experimental setting.

A more subtle source of BIAS lurks in the interaction between the experimenter and the participant. When an experimenter expects a certain response from participants in one condition, he or she may artificially generate or enhance this response by influencing these participants with (often unintentional) nonverbal cues, extra encouragement, or heightened attention. The bias that results from these behaviors is called the experimenter expectancy effect. Because participant and experimenter biases are systematically related to expectations of the experimental manipulation, they have the potential to confound the interpretation of experimental results. The same type of expectancy effects may lead the people handling the data to consciously or unconsciously skew the results in an expected direction.

The double-blind procedure, which is also referred to as a "masked condition" or an "experimentally naïve" procedure, is often used in medical research. For example, in a drug trial, a sugar pill identical in appearance to the experimental pill is administered to the control group. The experimental and sugar pill treatments are randomly assigned by an investigator who does not interact with participants. The identity of the pills keeps both the participants and the experimenters who actually administer the drugs blind to the nature of each experimental trial. Double blind procedure has the added benefit of preventing those administering the treatment from succumbing to the temptation to give needier patients the experimental drug rather than the placebo, thereby distorting the results.

Many social science investigations involve treatments that, unlike a pill, cannot be engineered to look the same to treatment and control subjects. Consider a study comparing the effects of telephone versus e-mail messages on attitude change; here, both participants and experimenters are fully aware of whether the study involves a phone call or an e-mail. In such instances, double blindness is approximated by keeping both participants and experimenters unaware of the hypotheses, or by minimizing contact between participants and the experimenter who knows the hypothesis. For example, a third party can administer the treatment before the experimenter collects data on the DEPENDENT VARIABLE. In sum, a "blind" PROTOCOL helps the investigator to minimize systematic human bias that is introduced when participants, experimenters, and data analysts are aware of the experimental treatment. Double-blind procedures afford the investigator more confidence when drawing causal relationships between the treatment and the outcome variable.

—Donald P. Green and Elizabeth Levy Paluck

REFERENCES

- Aronson, E., Carlsmith, M., & Ellsworth, P. C. (1990). *Methods of research in social psychology*. New York: McGraw-Hill.
- Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston: Allyn & Bacon.

DUAL SCALING

This is a method for multidimensional quantification of categorical data in a two-way table (e.g., a CONTINGENCY TABLE, an examinees-by-questions table of multiple-choice responses). The word "dual" reflects its symmetric scaling of rows and columns of a table. Dual scaling determines weights for options of multiple-choice items in such a way that scores of the subjects, given by weighted sums of weights of chosen options, attain greatest discriminability, hence, maximal internal consistency; in turn, those option weights are expressed as weighted sums of scores of the examinees who chose the options. This property is often referred to as Louis Guttman's principle of internal consistency. Frederic M. Lord verified that such a weighting scheme as this maximizes the internal consistency reliability of the derived scores.

HISTORICAL DEVELOPMENT

Early work on multidimensional quantification of categorical data can be traced back to Marion Richardson and G. Frederick Kuder in 1933, using what Paul Horst in 1935 called the method of reciprocal averages. H. O. Hirschfeld in 1935 presented a mathematically equivalent method, simultaneous linear regressions, as termed by James C. Lingoes in 1964. Ronald A. Fisher in 1940 proposed DISCRIMINANT ANALYSIS of categorical data, and K. Maung in 1941 showed that Fisher's quantification method could be equivalently formulated by four distinct objective functions. Whereas Maung's paper was for the contingency table, Guttman in 1941 presented three mathematically equivalent approaches to the quantification of multiple-choice data, also extending it to paired comparison and rank order data in 1946. The foundation for quantification was firmly laid by 1946. In 1950, Chikio Hayashi launched a project on Hayashi's quantification theory in Japan, Type III of which corresponds to dual scaling of the contingency table. In the early 1960s, Jean-Paul Benzécri started research in France, which was developed into CORRESPONDENCE ANALYSIS for the contingency table and multiple correspondence analysis for multiple-choice data. Both Hayashi and Benzécri had many followers, as referred to by the Hayashi School and the Benzécri School of data analysis. Similar consorted efforts led to the Leiden Group (headed by Jan de Leeuw) in the Netherlands and the Toronto Group (headed by Shizuhiko Nishisato) in Canada in the late 1960s.

A plethora of names have been proposed for the method, such as Guttman scaling (Guttman); the method of reciprocal averages (Richardson, Kuder, Horst); simultaneous linear regressions (Hirschfeld); Fisher's appropriate scoring (Fisher); Hayashi's theory of quantification Type III (Hayashi); optimal scaling (R. Darrell Bock); correspondence/multiple-correspondence analysis (Benzécri, Brigitte Escofier); biplot (Ruben Gabriel, John Gower, David Hand); homogeneity analysis (de Leeuw, Willem Heiser, Jacqueline Meulman); and dual scaling (Nishisato). The name *dual scaling* was proposed in 1976 in response to the criticism that the popular name *optimal scaling* was not specific enough to describe the method. According to Meulman in 1998, dual scaling is a "comprehensive framework for multidimensional analysis of categorical data," (p. 291) because it handles a wider range of categorical data than such methods as

correspondence analysis and optimal scaling, of which applications are mainly to incidence data.

BASIC FORMULAS

The method employs singular value decomposition for categorical data. Consider a two-way contingency table with the typical element being the frequency f_{ij} in row i and column j . The singular value decomposition of this element can be expressed as

$$f_{ij} = \frac{f_{i.}f_{.j}}{f_{..}}[1 + \rho_1 y_{i1}x_{j1} + \rho_2 y_{i2}x_{j2} + \cdots + \rho_k y_{ik}x_{jk} + \rho_K y_{iK}x_{jK}], \quad (1)$$

where $f_{i.}$ is the marginal frequency of row i , $f_{.j}$ is that of column j , ρ_k is the k th singular value ($\rho_1 \geq \rho_2 \geq \cdots \geq \rho_K$), y_{ik} is the optimal weight of row i of the k th solution, and x_{jk} is that of column j . The solution, corresponding to 1 inside the bracket, is called a trivial solution, typically deleted from analysis, and is the frequency expected when the rows and the columns are statistically independent. The others are called proper solutions. For any proper solution, say, solution k , there exist dual relations (Nishisato, 1980) or transition formulas (Benzécri et al. in 1973): For an $m \times n$ data matrix,

$$y_{ik} = \frac{1}{\rho_k} \frac{\sum_{j=1}^n f_{ij}x_{jk}}{f_{i.}} \quad \text{and} \quad x_{jk} = \frac{1}{\rho_k} \frac{\sum_{i=1}^m f_{ij}y_{ik}}{f_{.j}}. \quad (2)$$

The equations are nothing but an expression of simultaneous linear regressions of rows on columns and columns on rows. It also indicates the basis for the method of reciprocal averages: Start with arbitrary row weights to calculate weighted column means, which are, in turn, used as weights to calculate weighted row means, and continue the process. This reciprocal averaging scheme always converges (Nishisato, 1980) to the optimal sets of row weights and column weights, the proportionality constant being the singular value. Nishisato (1980) calls weights y and x *normed weights* and ρy and ρx *projected weights*, also referred to as *standard coordinates* and *principal coordinates*, respectively (see Greenacre, 1984). Geometrically, y and ρx span the same space, and so do ρy and x . The angle of the discrepancy between row space and column space is given by the cosine of the singular value (Nishisato & Clavel, 2003).

Table 1

Party	Plan	Ranking						Dominance Matrix					
		(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
Subject	1	6	1	5	4	3	2	-5	5	-3	-1	1	3
	2	6	1	5	2	4	3	-5	5	-3	3	-1	1
	3	3	5	2	4	1	6	1	-3	3	-1	5	-5
	4	3	4	2	6	1	5	1	-1	3	-5	5	-3
	5	5	3	1	4	6	2	-3	1	5	-1	-5	3
	6	2	6	3	5	4	1	3	-5	1	-3	-1	5
	7	1	2	4	5	3	6	5	3	-1	-3	1	-5
	8	4	3	2	6	5	1	-1	1	3	-5	-3	5
	9	2	1	4	5	3	6	3	5	-1	-3	1	-5
	10	6	1	4	3	5	2	-5	5	-1	1	-3	3

Contingency tables and multiple-choice data, expressed as response patterns, are examples of *incidence data*, which are decomposed into the expression of equation (1). In contrast, *dominance data* include rank-order data and paired comparison data, for which the observations are first coded as follows:

$${}_i f_{jk} = \begin{cases} 1 & \text{if Subject } i \text{ prefers Object } j \text{ to } k, \\ 0 & \text{if Subject } i \text{ makes a tied judgment,} \\ -1 & \text{if Subject } i \text{ prefers Object } k \text{ to } j. \end{cases} \quad (3)$$

These are transformed to dominance numbers by

$$e_{ij} = \sum_{j=1}^n \sum_{k=1, (j \neq k)}^n {}_i f_{jk}. \quad (4)$$

The matrix of dominance numbers has the property that row elements sum to zero. Thus, to arrive at weights for subjects and weights for objects, the method of reciprocal averages, for example, is carried out with the understanding that each cell of the dominance matrix is based on $n - 1$ responses.

INCIDENCE DATA

In 1993, Nishisato classified categorical data into two types, incidence data and dominance data. Incidence data consist of 1, 0, or frequencies; a trivial solution exists; the chi-square metric is used; and even when all variables are perfectly correlated, more than one dimension is generally needed for a complete description of data. Incidence data are the data type for

which most quantification methods mentioned earlier are developed. See Gifi (1990), Greenacre (1984), and Nishisato (1980, 1994) for analytical examples of such incidence data as contingency tables, multiple-choice data, and sorting data.

DOMINANCE DATA

Dominance data consist of responses *greater*, *equal*, or *smaller*; no trivial solution is involved; the Euclidean metric is used; and when variables are perfectly correlated, a single dimension is sufficient to describe data. Because dual scaling also handles dominance data, let us look at one example. Ten subjects ranked the following six Christmas party plans according to the order of their preference: (1) potluck in the group during the daytime, (2) pub-restaurant crawl after work, (3) reasonably priced lunch in an area restaurant, (4) evening banquet at a hotel, (5) potluck at someone's home after work, (6) ritzy lunch at a good restaurant. The data from *Psychometrika* (Nishisato, 1996) and the corresponding dominance matrix are in Table 1. Relative variances of five proper solutions are 0.45, 0.27, 0.14, 0.08, and 0.06, respectively. Look at the first two proper solutions, and a joint plot of the *normed* weights y of subjects and the *projected* weights ρx of the party plans, the practice that leads to a solution to Clyde H. Coombs's problem of multidimensional unfolding (Nishisato, 1994) (see Figure 1). From this graph, calculate the distances between subjects and party plans, and rank the distances within each subject, to obtain the rank-2 approximation to the input ranks (see Table 2). Look at Subject 2: The ranking of the distances reproduces exactly one's ranking of the plans.

Table 2

	Squared Distances						Rank-2 Approximation					
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
S1	1.94	0.73	1.72	1.06	1.67	1.37	6	1	5	2	4	3
S2	2.20	0.98	2.02	1.36	1.91	1.71	6	1	5	2	4	3
S3	0.67	1.75	1.21	1.53	0.81	1.98	1	5	3	4	2	6
S4	0.49	1.63	1.01	1.38	0.69	1.79	1	5	3	4	2	6
S5	1.72	1.58	1.18	1.27	1.77	0.54	5	4	2	3	6	1
S6	1.70	2.38	1.37	1.95	1.96	1.51	3	6	1	4	5	2
S7	1.10	1.42	1.55	1.45	0.89	2.14	2	3	5	4	1	6
S8	1.61	1.70	1.07	1.33	1.71	0.76	4	5	2	3	6	1
S9	1.40	1.28	1.74	1.45	1.11	2.18	3	2	5	4	1	6
S10	2.13	1.02	1.79	1.21	1.91	1.25	6	1	4	2	5	3

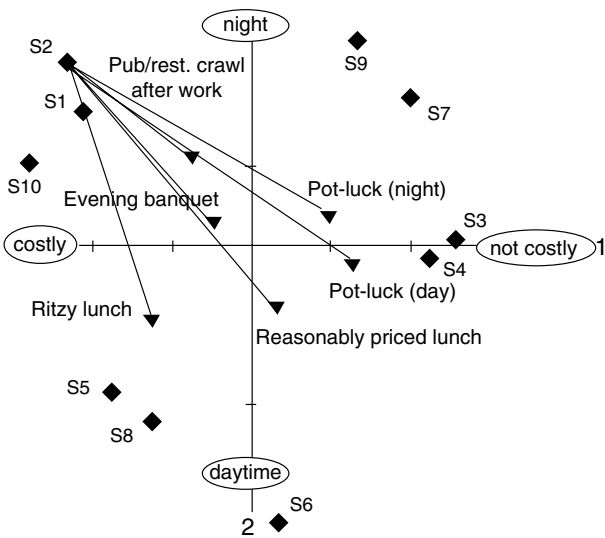


Figure 1 Joint Plot of Respondents and Party Plans

Other cases require more dimensions than two. Overall, the rank-2 approximation is close to the rank-5 data. The graph reveals individual differences in ranking as affected by the party cost and the party times. Observe their closeness to particular plans.

—Shizuhiko Nishisato

REFERENCES

Benzécri, J. P., et al. (1973). *L'analyse des données: II. L'analyse des correspondances*. Paris: Dunod.
Gifi, A. (1990). *Nonlinear multivariate analysis*. New York: Wiley.

Greenacre, M. J. (1984). *Theory and method of correspondence analysis*. London: Academic Press.
Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive statistical analysis*. New York: Wiley.
Meulman, J. J. (1998). Book review of W. J. Krzanowski, & F. H. C. Marriott, *Multivariate analysis. Part I. Distributions, ordinations, and inference*, *Journal of Classification*, 15, 287–293.
Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.
Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, NJ: Lawrence Erlbaum.
Nishisato, S. (1996). Gleaning in the field of dual scaling. *Psychometrika*, 61, 559–599.
Nishisato, S., & Clavel, J.G. (2003). A note on between-set distances in dual scaling and correspondence analysis. *Behaviormetrika*, 30, 87–98.

DUMMY VARIABLE

A set of dummy variables allows us to specify categorical explanatory variables as independent variables in statistical analyses. Some of the information we wish to incorporate into statistical models is organized as classification schemes that represent a set of nominal categories. Marital status, gender, race/ethnicity, political party, or industrial sector can all be represented as a set of exhaustive and mutually exclusive categories, with each observation being assigned to a single category. In the simplest case, we have two categories (e.g., men and women), which we can represent with a single dummy variable coded “1” if a man, “0” if not a

man. This type of CODING is called binary coding and indicates the presence of an attribute or membership in a particular category. For classifications with more than two categories, we can fully represent the categorical information and produce coefficients that provide interpretable information if we strategically code $j - 1$ variables (Hardy, 1993; Hardy & Reynolds, 2004).

When specified on the right-hand side of an equation (as independent variables), researchers may want to know how differences in marital status, for example, are related to variation in a dependent variable, such as depressive symptoms. At issue is how to represent the information contained in a j -category classification scheme without imposing unrealistic distributional assumptions. If, for example, marital status was reported for a sample of adult men and women in the United States, we could specify categories such as (1) currently married, (2) currently divorced, (3) currently separated, (4) currently widowed, and (5) never married. If we simply included the variable "marital status" in a statistical model, without modifying in any way the coding of 1 through 5 assigned to the different categories, we would be assuming interval-level measurement of marital status, that is, that the numerical difference of one unit between "currently married" and "currently divorced" is equivalent to the unit difference between "currently widowed" and "never married," but only half the difference between "currently widowed" and "currently divorced," and that these categories were appropriately ordered, with "currently married" being scored at the lowest value in the range and "never married" at the highest value. This approach is clearly nonsensical; the numbers assigned to categories as well as the order of the categories are arbitrary. We require an approach consistent with nominal distributional properties, which, in this case, suggests that we need a numerical way to isolate the distinctive categories so that we can rely on counts (i.e., frequencies) and sample partitioning to provide appropriate estimates.

Although coding possibilities are many, three approaches are most frequently used: BINARY CODING, CONTRAST CODING, and EFFECTS CODING. To return to the example of marital status, we require four dummy variables to fully represent the information in the five-category classification. Using binary coding and choosing "currently married" as our reference group (the group to which remaining categories will be compared), we can define four dummy variables named "divorced," "separated," "widowed," and

"never." Respondents whose marital status matches the variable name are coded "1" on that dummy variable. For example, currently divorced respondents are coded "1" on "divorced" and coded "0" on the remaining three variables; respondents who were never married are coded "1" on "never," "0" otherwise, and so on. That four dummy variables fully capture the information of a five-category nominal variable is demonstrated by the fact that any respondent coded "0" on all four variables must be married.

Regardless of which coding scheme is used, the appropriate *inference* test to determine whether the classification variable—in this example, marital status—improves the fit of the model is the F -test comparing model fit with and without the set of dummy variables. Dummy variables expand the flexibility of the GENERAL LINEAR MODEL, allowing us to test a broader range of hypotheses, including direct and indirect effects, INTERACTIONS, and different functional forms.

Binary-coded dummy variables also can be specified as dependent variables in statistical models. A set of estimation techniques for models with limited dependent variables includes binary probit and binary LOGIT models (Long & Cheng, 2004). Discrete time EVENT HISTORY models also use binary coded dependent variables to indicate whether or not an event has occurred (Allison, 2004). Binary-coded dependent variables are also used in various selection models, such as the Heckman two-stage procedure and endogenous switching models (Fu, Mare, & Winship, 2004).

—Melissa A. Hardy

REFERENCES

- Allison, P. D. (2004). Event history analysis. In M. A. Hardy & A. Bryman (Eds.), *Handbook of data analysis*. London: Sage.
- Fu, V., Mare, R., & Winship, C. (2004). Sample selection bias models. In M. A. Hardy & A. Bryman (Eds.), *Handbook of data analysis*. London: Sage.
- Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage.
- Hardy, M. A., & Reynolds, J. (2004). Incorporating categorical information into regression models: the utility of dummy variables. In M. A. Hardy & A. Bryman (Eds.), *Handbook of data analysis*. London: Sage.
- Long, J. S., & Cheng, S. (2004). Regression models for categorical outcomes. In M. A. Hardy & A. Bryman (Eds.), *Handbook of data analysis*. London: Sage.

DURATION ANALYSIS. See EVENT HISTORY ANALYSIS

DURBIN-WATSON STATISTIC

The Durbin-Watson (1950) statistic (DW or d) is a commonly used and routinely reported diagnostic test for the presence of first-order auto or SERIAL CORRELATION in the error of a TIME-SERIES REGRESSION model. The statistic is calculated using the RESIDUALS from the regression we care about, $\hat{\mu}_t$ as follows:

$$DW = \frac{\sum_2^n (\hat{\mu}_t - \hat{\mu}_{t-1})^2}{\sum_1^n \hat{\mu}_t^2}.$$

The NULL HYPOTHESIS for the test is that there is no first-order autocorrelation in the regression ERRORS. The test has two alternatives: positive and negative first-order autocorrelation.

It is illustrative to express the test statistic directly in terms of the first-order autocorrelation ρ :

$$DW \approx 2(1 - \rho).$$

Written in this form, it is easy to see the relationship between the value of the test statistic and the degree of autocorrelation in the model residuals. Specifically, if there is no autocorrelation ($\rho = 0$), DW is approximately 2; if there is perfect positive autocorrelation ($\rho = 1$), DW is approximately 4; and if there is perfect negative autocorrelation ($\rho = -1$), DW is approximately 0.

The CRITICAL VALUES for the test vary as a function of SAMPLE size and the DEGREES OF FREEDOM associated with the regression of interest. Typically, the DW test is computed for the alternative hypothesis of positive first-order autocorrelation so that we are looking for a value of DW that is significantly smaller than 2. Given difficulties in deriving the distribution of the DW statistic under the null hypothesis, two critical values are needed for each alternative hypothesis.

The test is appropriate only if there are no LAGGED dependent variables on the right-hand side of the regression we care about. If the regression contains a lagged dependent variable, alternative tests include Durbin's alternative, Durbin's H, or Lagrange multiplier tests. Similarly, DW will not detect higher orders

of autocorrelation. To test for second and higher orders of autocorrelation in the errors, one can use Lagrange multiplier tests.

Finding a significant DW statistic indicates autocorrelation, which will invalidate HYPOTHESIS tests on the regression PARAMETERS we care about. A significant DW statistic is generally evidence of MODEL misspecification, but may indicate that the errors are themselves serially correlated (Wooldridge, 2000). The appropriate remedial response varies based on the cause of the problem. The model may require that the right-hand side of the equation include a lagged dependent variable or other dynamic REGRESSORS. Alternatively, we may need to estimate the regression with feasible GENERALIZED LEAST SQUARES (GLS) by transforming the variables, reestimating the original regression to remove the autocorrelation, and estimating the resulting regression.

To illustrate, suppose we wish to model the voter turnout rate of women in the United States, (v_t), from 1921 to the present, as a function of the mean level of education of women (e_t) and percent of women employed outside the home (w_t). To test for first-order serial correlation, we would first save the residuals from the following regression model:

$$v_t = \beta_0 + \beta_1 e_t + \beta_2 w_t + \mu_t.$$

The residuals would then be used to compute the DW statistic as above. For a regression with a sample of size $t = 45$ and 3 degrees of freedom (two regressors plus a constant), the 5% critical values for the alternative of positive autocorrelation are 1.45 (lower) and 1.62 (upper). Assuming an estimated $DW = 1.8$, we could not reject the null hypothesis that there is no positive autocorrelation, and we would infer that there is probably no first-order autocorrelation. If the estimated $DW = 1.4$, we would reject the null hypothesis of no positive autocorrelation. For all values between 1.45 and 1.62, the test is inconclusive, and we would infer that there probably is first-order autocorrelation.

—Suzanna De Boef

REFERENCES

- Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regressions I. *Biometrika*, 37, 409–428.
- Wooldridge, J. M. (2000). *Introductory econometrics: A modern approach*. Cincinnati, OH: South-Western Thomson Learning.

DYADIC ANALYSIS

Often in social research, pairs of people (e.g., married couples, roommates, friends, and coworkers) are measured. The statistical analysis of data from pairs is called *dyadic analysis*. The analysis of dyadic data is much more complicated than the analysis of data from individuals (Kashy & Kenny, 2000). Particular consideration needs to be given to the interdependence of the data and to the UNIT OF ANALYSIS.

Typically, the scores of the two people are correlated, resulting in what is called *interdependence*. The correlation of scores invalidates the use of SIGNIFICANCE TESTING from an analysis that uses person as the unit of analysis. The *p* values from testing may be larger or smaller than they appear to be. This interdependence of scores can even be negative, such that the two members of the dyad are more different from one another than two randomly chosen people. The degree of interdependence can be measured by a Pearson correlation coefficient when members can be distinguished by some variable. For example, heterosexual married couples can be distinguished by their gender. When members cannot be distinguished (e.g., coworkers), the INTRACCLASS CORRELATION should be used to test for measuring and testing nonindependence.

To solve the problems created by nonindependence, one typically needs to make dyad, not person, the unit of analysis. In some cases, one can conduct analyses using the sum and/or the difference between the two members of the dyad as the variables in the analysis.

Although nonindependence can be viewed as a problem, it represents an opportunity to study various dyadic phenomena. For instance, reciprocity, compensation, synchrony, and similarity imply nonindependence in the responses of the members of the dyad. Sources of nonindependence such as these can be due to nonrandom pairing (e.g., assortative mating), common fate, and mutual influence.

Occasionally in dyadic analysis, an index is created for each dyad. For example, the similarity of the members' leisure preferences is indexed by a correlation coefficient. Although dyad indexes can be informative, there are several pitfalls in their interpretation (Cronbach, 1955).

It can be useful in dyadic analysis to study how it is that the characteristics of one person influence the behavior of the person and the person's partner. So, for each person, there is an outcome that is regressed on the person's own characteristics and the characteristics of the person's partner. As an example, it could be examined how the level of depression of the person and the person's roommate affects the person's satisfaction with the roommate. Kashy and Kenny (2000) refer to this model as the Actor-Partner Interdependence Model. Alternatively, it is possible to estimate models of mutual influence using feedback loops and instrumental variables (Duncan, Haller, & Portes, 1975).

Sometimes, with dyadic data, each person is paired with several others. For instance, each person might rate everyone else in her sorority. Such round-robin data requires the use of the SOCIAL RELATIONS MODEL.

—David A. Kenny

REFERENCES

- Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin*, 52, 177–193.
- Duncan, O. D., Haller, A., & Portes, A. (1968). Peer influence on aspiration: A reinterpretation. *American Journal of Sociology*, 75, 119–137.
- Kashy, D. A., & Kenny, D. A. (2000). The analysis of data from dyads and groups. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 451–477). New York: Cambridge University Press.

DYNAMIC MODELING. See SIMULATION

E

ECOLOGICAL FALLACY

In 19th-century Europe, suicide rates were higher in countries that were more heavily Protestant; Emile Durkheim inferred that suicide was promoted by the social conditions of Protestantism (*Le suicide*, 1897). This is an “ecological inference”—a conclusion about individual behavior drawn from data about aggregate behavior.

The Protestant countries, of course, were different from the Catholic countries in many ways besides religion (the problem of “confounding”). Moreover, Durkheim’s data did not tie individual suicides to any particular religious faith: He had rates for countries, not for religious denominations within geographic areas.

The problem of confounding must be dealt with in any observational study, but the second problem is specific to ecological studies: Putative causes and effects are measured for groups rather than for individuals. Moreover, the analytic interest is in one kind of grouping—for Durkheim, religion—whereas data are available for an entirely different sort of grouping (geography).

If there is no confounding, the expected difference between effects for groups and for individuals is “aggregation bias”; in general, the difference is attributable partly to confounding and partly to aggregation bias. The “ecological fallacy” consists of thinking that relationships observed for groups necessarily hold for individuals: If countries with more Protestants tend to have higher suicide rates, then Protestants must be more likely to commit suicide.

Ecological studies can provide useful clues, but conclusions about individuals are in general only weakly supported by data on groups. The source of the problem is confounding and aggregation bias. Indeed, as shown by Robinson (1950), individual-level relationships can be reversed by aggregation. This is “Simpson’s Paradox” for the correlation coefficient.

Statistical procedures have been proposed for disentangling individual-level from group-level behavior, including “ecological regression” and “cross-level” or “hierarchical” regression models. However, each method makes its own rather strong behavioral assumptions, which may seem implausible when stated explicitly. For instance, ecological regression makes the “constancy assumption.” According to this assumption, with an application like Durkheim’s, individual behavior depends on religious affiliation but not geographical location. Protestants all over Europe must have similar propensities to commit suicide, and Catholics are just as homogeneous.

Generally, the identifying assumptions in the models cannot be validated by the data. Moreover, in test situations where individual-level data are available—so that estimates can be compared to reality—the track record of the models is mixed at best. For additional discussion from various perspectives, see the references below.

—David A. Freedman

REFERENCES

- Firebaugh, G. (2001). Ecological fallacy. *International encyclopedia for the social and behavioral sciences* (Vol. 6, pp. 4023–4026). Oxford, UK: Pergamon.

- Freedman, D. A. (2001). Ecological inference and the ecological fallacy. *International encyclopedia for the social and behavioral sciences* (Vol. 6, pp. 4027–4030). Oxford, UK: Pergamon.
- Freedman, D. A., Klein, S. P., Ostland, M., & Roberts, M. R. (1998). Review of "A solution to the ecological inference problem." *Journal of the American Statistical Association*, 93, 1518–1522. (Discussion appears in Vol. 94, pp. 352–357.)
- Goodman, L. (1953). Ecological regression and the behavior of individuals. *American Sociological Review*, 18, 663–664.
- Grofman, B., & Davidson, C. (1992). *Controversies in minority voting: The voting rights act in perspective*. Washington, DC: Brookings Institution.
- King, G. (1997). *A solution to the ecological inference problem*. Princeton, NJ: Princeton University Press.
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351–357.

ECOLOGICAL VALIDITY

Ecological validity refers to the extent to which behavior indicative of behavior studied in one environment (often, reference is to a laboratory setting) can be taken as characteristic of (or generalizable to) an individual's cognitive processes in a range of other environments (often glossed as "everyday" or "natural").

Discussions of the problem of ecological validity first came to prominence in psychological research in the United States owing to the work of Egon Brunswik (1943) and Kurt Lewin (1943), two German scholars who emigrated to the United States in the 1930s. Following in this tradition, developmental psychologist Urie Bronfenbrenner argued that there are three conditions that ecologically valid research must fulfill: It must (a) maintain the integrity of the real-life situations it is designed to investigate, (b) be faithful to the larger social and cultural contexts from which the subjects come, and (c) be able to demonstrate that the experimental manipulations and outcomes are "perceived by the participants in a manner consistent with the conceptual definitions explicit and implicit in the research design" (Bronfenbrenner, 1979, p. 35). In similar terms, Neisser (1976) noted marked discontinuities between the spatial, the temporal, and the intermodal relationships between real objects and events, on one hand, and the objects and

events characteristic of laboratory-based research, on the other, as a fundamental shortcoming of cognitive psychology.

In many discussions of ecological validity, it is assumed that it is possible to discover and directly observe the ways that laboratory tasks occur (or don't occur) in nonlaboratory settings (e.g., Simon, 1976). A variety of findings in contemporary research, however, indicate that the requirements for establishing ecological validity place an enormous analytical burden on social scientists (see Cole, 1996, chaps. 8–9 for an extended treatment of the associated issues). Once we move beyond the laboratory in search of real-world analogues, the ability to identify tasks is markedly weakened. Failure to define the parameters of the analyst's task or failure to ensure that the task-as-discovered is also the subject's task vitiate the enterprise.

The problem arises, as Valsiner and Benigni (1986) point out, because standardized cognitive experimental procedures are meant to embody closed analytic systems. Consequently, attempting to establish task equivalence in order to generalize beyond the experimental circumstances amounts to imposing a closed system upon a more open behavioral system. Insofar as the social scientist's closed system does not capture veridically the elements of the open system it is presumed to model, experimental results systematically misrepresent the life process from which they are derived.

—Michael Cole

REFERENCES

- Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.
- Brunswik, E. (1943). Organismic achievement and environmental probability. *The Psychological Review*, 50, 255–272.
- Cole, M. (1996). *Cultural psychology: A once and future discipline*. Cambridge, MA: Harvard University Press.
- Lewin, K. (1943). Defining the "field at a given time." *Psychological Review*, 50, 292–310.
- Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. San Francisco: W. H. Freeman.
- Simon, H. A. (1976). Discussion: Cognition and social behavior. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior*. Hillsdale, NJ: Erlbaum.
- Valsiner, J., & Benigni, L. (1986). Naturalistic research and ecological thinking in the study of child development. *Developmental Review*, 6(3), 203–223.

ECONOMETRICS

Literally interpreted, *econometrics* means “economic measurement.” Although measurement is an important part of econometrics, the scope of econometrics is much broader, as can be seen from the following quotations.

Econometrics, the result of a certain outlook on the role of economics, consists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results. (Tintner, 1968, p. 74)

Econometrics may be defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the analysis of economic phenomena. (Goldberger, 1964, p. 1)

The method of econometric research aims, essentially, at a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier. (Haavelmo, 1944, p. iii)

As the preceding definitions suggest, econometrics is an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics. Instead of being an adjunct to one of these disciplines, econometrics is now taught as a separate subject in most undergraduate and graduate departments of business and economics, for reasons explained below.

Economic theory makes statements or hypotheses that are mostly qualitative in nature. For example, microeconomic theory states that, other things remaining the same (the famous *ceteris paribus* clause), a reduction in the price of a commodity is expected to increase the quantity demanded of that commodity. Thus, economic theory postulates a negative or inverse relationship between the price and the quantity demanded of a commodity. The theory itself, however, does not provide any numerical measure of the relationship between the two. That is, it does not tell by how much the quantity demanded will go up or down as a result of a certain change in the price of that commodity. It is the job of the econometrician to provide such numerical estimates. Put

differently, econometrics gives empirical content to economic theory. Sometimes econometrics can refine an existing theory by explicitly considering variables that were subsumed in the *ceteris paribus* clause. Qualitative variables such as sex, religion, and ethnicity have been added to labor market studies to explain labor force participation (i.e., the decision to work or not). Traditionally, variables such as hourly earnings and level of education were considered the primary determinants of labor force participation.

The main concern of mathematical economics is to express economic theory in mathematical form (equations) without regard to measurability or empirical verification of the theory. The econometrician uses the mathematical equations provided by the mathematical economist but puts these equations in such a form that they lend themselves to empirical testing. This conversion of mathematical equations into econometric equations (i.e., STOCHASTIC equations) requires a great deal of ingenuity and practical skill.

Economic statistics is concerned primarily with collecting, processing, and presenting economic data in the form of charts and tables. These are the jobs of the economic statistician. It is he or she who is primarily responsible for collecting data on such variables as gross domestic product (GDP), employment, unemployment, and prices. The data thus collected constitute the raw data for econometric work. The economic statistician does not go any further, not being concerned with using the collected data to test economic theories. Someone who does that becomes an econometrician.

Although mathematical statistics provides many tools used in the trade, the econometrician needs special methods in view of the unique nature of most economic data, namely, that the data are not generated as the result of controlled experiments. The econometrician, like the meteorologist, generally depends on data that cannot be controlled directly. As Spanos (1999) observed:

In econometrics the modeler is often faced with *observational* as opposed to *experimental* data. This has two important implications for empirical modeling in econometrics. First, the modeler is required to master very different skills than those needed to analyze experimental data. . . . Second, the separation of the data collector and data analyst requires the modeler to familiarize himself/herself thoroughly with the nature and structure of data in question. (p. 21)

METHODOLOGY OF ECONOMETRICS

How do econometricians proceed in their analysis of an economic problem? That is, what is their methodology? Broadly speaking, there are three approaches to econometric methodology: classical, Leamer's (1978), and Hendry's (1995).

A discussion of all these approaches to econometric methodology would go far afield. The main features of the classical methodology are presented here because of its historical importance and the fact that this methodology still dominates empirical work in many areas. After discussion of this methodology, the other two approaches will be discussed.

Classical Methodology

Classical methodology proceeds as follows:

Economic theory → Mathematical model of theory → Econometric model of theory → Data → Estimation of econometric model → Hypothesis testing → Forecasting or prediction → Using the model for control or policy purposes.

Each arrow indicates a sequential or stepwise process.

To best illustrate the classical methodology, we consider the well-known Keynesian theory of consumption.

Statement of Theory or Hypothesis

Keynes (1936) stated that

The fundamental psychological law ... is that men [and women] are disposed, as a rule and on average, to increase their consumption as their income increases, but not as much as the increase in their income. (p. 36)

In short, Keynes postulated that the *marginal propensity to consume* (MPC), the rate of change of consumption for a unit (say, a dollar) change in income, is greater than zero but less than 1.

Specification of the Mathematical Model of Consumption

Although Keynes postulated a positive relationship between consumption and income, he did not specify the precise form of the functional relationship between the two. For simplicity, a mathematical economist

might suggest the following form of the Keynesian consumption function:

$$Y = \beta_1 + \beta_2 X; \quad 0 < \beta_2 < 1 \quad (1)$$

where Y is consumption expenditure and X is income. β_1 and β_2 , the parameters of the model, are, respectively, the intercept and slope coefficients, with the slope coefficient representing the MPC.

Specification of the Econometric Model of Consumption

The mathematical model of the consumption function given in equation (1) is of limited use to the econometrician because it assumes that there is an *exact* or *deterministic* relationship between consumption expenditure and income. But relationships between economic variables generally are inexact. To allow for the inexact relationship, the econometrician would modify the mathematical model of the consumption function by writing it as

$$Y = \beta_1 + \beta_2 X + u \quad (2)$$

where u , known as the *disturbance*, or ERROR term, is a RANDOM (or stochastic) VARIABLE that has well-defined properties. The error term u may represent all those factors that affect consumption but are not explicitly accounted for in equation (2). Equation (2) is an example of an *econometric model*. More technically, it is an example of a LINEAR REGRESSION MODEL.

Obtaining Data

To estimate equation (2)—that is, to obtain the numerical estimates of β_1 and β_2 , which in the present context are known as regression coefficients—we need data. For illustrative purposes, we obtained data on Y (personal consumption expenditure) and X (GDP) for the United States for the period 1982–1996, the data being measured in 1992 in billions of dollars.

Estimation of Econometric Model

Using the method of ORDINARY LEAST SQUARES (OLS), and based on the U.S. data, the following regression results were obtained:

$$\begin{aligned} \hat{Y}_t &= -184.0780 + 0.7064X_t \\ SE : & \quad (46.2619) \quad (0.0078) \\ R^2 &= \quad 0.9984 \end{aligned} \quad (3)$$

where the hat over Y_t means that it is the estimated value of mean, or average, consumption expenditure, and where figures in the parentheses are the estimated standard errors. The subscript t indicates time. [The classical linear model makes certain assumptions about the error term u , namely, that it has a mean of zero, that it has a constant variance, and that the errors are uncorrelated. For hypothesis testing, it is often assumed that the error term is normally distributed. If the normality assumption is invoked, we can use the METHOD OF MAXIMUM LIKELIHOOD ESTIMATION (ML) to estimate the parameters of the regression model. The ML estimators of β_1 and β_2 are the same as those obtained by OLS.]

Hypothesis Testing

Keynes asserted that the marginal propensity to consume is positive but less than 1. In the estimated regression model given in equation (3), the MPC is 0.7064. Is it statistically different from 1? In other words, we wish to test the null hypothesis that the true MPC is 1. How does one test this hypothesis?

Given the assumptions of the CLASSICAL LINEAR REGRESSION model, one can test the hypothesis by the T -TEST, which in the present case gives

$$t = \frac{0.7064 - 1}{0.0078} = -3.7641. \quad (4)$$

Because there are 15 observations in the data, this t value has 13 DEGREES OF FREEDOM. Therefore, this t value is significant at the 1% level of significance (the actual p value is much lower). We can thus reject the null hypothesis that the true MPC is 1, perhaps substantiating Keynes' hypothesis.

Forecasting or Prediction

Assuming that the econometric model is a reasonably good approximation of reality, one can use it to forecast the future value of Y on the basis of a known or expected future value of the predictor variable X . Suppose we were to predict the mean consumption expenditure for 1997. Suppose that the value of GDP for 1997 was \$7,269.8 billion (which was the actual value). Using this value of GDP in equation (3), the reader can verify that the predicted mean consumption expenditure is about \$4,951.31 billion.

Use of Models for Control or Policy Purposes

Suppose that the government believes that the consumption expenditure of about \$4,900 billion will keep

unemployment at its current level. What level of GDP will guarantee the target level of consumption expenditure? Assuming that the estimated consumption expenditure regression in equation (3) is reasonably good, all one has to do is to put \$4,900 on the left-hand side of that equation and solve for X . Simple arithmetic will show that an expenditure level of \$7,200 billion will maintain the current level of unemployment. In this exercise, Y is the *target variable* and X is the *control variable*.

The preceding eight-step procedure is, in a nutshell, the classical methodology. There are several critics of this methodology. Two of them are considered below.

Hendry's Top-Down Methodology

According to Hendry (1995), the classical methodology is too static. That is, it does not take into account changes over time. In other words, he would develop econometric models in a dynamic setting. He would make his econometric model as comprehensive as possible and then whittle it down to a smaller model after considerable diagnostic testing. Because the bulk of Hendry's initial work was done at the London School of Economics (LSE), his methodology is also known as the LSE methodology.

The LSE starting point is that economic theory postulates a long-run, or equilibrium, relationship between economic variables, say Y ("permanent" consumption) and X ("permanent" income), the term "permanent" indicating the level that households regard as likely to persist in the future. This relationship can be summarized as

$$Y_t = \alpha X_t. \quad (5)$$

Hendry and his colleagues at LSE index the observations with the time subscript t because their methodology was developed to deal mainly with TIME-SERIES economic data.

Of course, the long-term relationship postulated in equation (5) takes time to achieve; therefore, the LSE methodology proposes the following type of dynamic procedure in order to reach equation (5):

$$Y_t = \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_m X_{t-m} + \delta_1 Y_{t-1} + \delta_2 Y_{t-2} + \cdots + \delta_l Y_{t-l} + u_t. \quad (6)$$

In this model, Y depends not only on current and lagged X values but also on its own lagged values, the lags being represented by the lagged subscript notation. Of course, more X variables and their lags can be introduced.

As equation (6) suggests, we regress Y at time t on the values of X at times t , $(t - 1)$, $(t - 2)$, $\dots$, and $(t - m)$ as well as the lagged values of Y at times $(t - 1)$, $(t - 2)$, $\dots$, $(t - l)$, the choice of the lagged terms m and l being an empirical question.

The model in equation (6) is an example of an *autoregressive distributed lag* (ADL) model. It is autoregressive because we are regressing the current Y value on its previous lagged values, and it is distributed because the effect of the X is spread over time. Such ADL models are also known as *dynamic models* because they explicitly consider the behavior of the Y variable over time. Once equation (6) is estimated, one can get back to the long-run relationship with appropriate algebraic manipulations.

Leamer's Methodology

Leamer is a strong critic of the classical econometric methodology, which he calls *average economic regression* (AER). In his view, in nonexperimentally collected data, the AER strategy is questionable. To him and other critics, once a model is specified, estimating its parameters and engaging in hypothesis testing is trivial, but the task of determining what the appropriate model is to begin with is very demanding. This task is the subject of *specimetrics*. According to Leamer (1978):

Specimetrics describes the process by which a researcher is led to choose one specification of the model rather than another; furthermore, it attempts to identify the inferences that may be properly drawn from a data set when the data generating mechanism is ambiguous. (p. v)

In short, Leamer and his followers strongly believe that one must pay very careful attention to specimetrics, that is, the choice of the appropriate model. Once that is done, one may then follow the classical or AER methodology.

Besides this fundamental difference between Leamer's and the AER methodologies, Leamer is a strong proponent of BAYESIAN METHODS. The classical and Hendry's approaches to econometric modeling are based on classical statistics. In a Bayesian approach to, say, regression models, "parameters" such as β_1 and β_2 in equation (1) are regarded as "random," with some probability distribution for each (called the *priors* in Bayesian language). Priors typically are chosen subjectively based on such factors as previous research and one's understanding of a phenomenon. Given the

data at hand, we estimate β_1 and β_2 and then, using the sample estimates and the priors, we obtain what are called *posterior* (probability) distributions of β_1 and β_2 . Thus, in the Bayesian approach, β_1 and β_2 are estimated partially from the data and partially by the prior information about these coefficients. This gives us the posterior (probability) distributions of these coefficients. In the next round of analysis, these posterior distributions become the prior distributions. Thus, the Bayesian approach is a "learning by doing" procedure. As more data become available, researchers revise prior knowledge about the coefficients.

In the classical regression procedure, on the other hand, β_1 and β_2 are regarded as fixed numbers, although their values are unknown. Given the data, one can use OLS or maximum likelihood (ML) to estimate the parameters and engage in hypothesis testing. If we have a new set of data, we start fresh, without incorporating the results from the previous sample. Thus, there is a fundamental difference between the two approaches.

—Damodar N. Gujarati

REFERENCES

- Goldberger, A. S. (1964). *Econometric theory*. New York: John Wiley & Sons.
- Gujarati, D. (2002). *Basic econometrics* (4th ed.). New York: McGraw-Hill.
- Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, 12 (Suppl.), preface, p. iii.
- Hendry, D. F. (1995). *Dynamic econometrics*. New York: Oxford University Press.
- Keynes, J. M. (1936). *The general theory of employment, interest, and money*. New York: Harcourt Brace Jovanovich.
- Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York: John Wiley & Sons.
- Mittelhammer, R. C., Judge, G. G., & Miller, D. J. (2000). *Econometric foundations*. New York: Cambridge University Press.
- Spanos, A. (1999). *Probability and statistical inference: Econometric modeling with observational data*. Cambridge, UK: Cambridge University Press.
- Tintner, G. (1968). *Methodology and mathematical economics and econometrics*. Chicago: University of Chicago Press.

EFFECT SIZE

A standardized numerical INDEX of the magnitude of an effect or relationship is called an *effect size*. The terms "magnitude" and "standardized" are particularly important in the above definition and deserve further

explanation. The magnitude of an effect is critical for the proper interpretation of social scientific findings. The alternative to consideration of effect size is to base conclusions solely on the results of statistical SIGNIFICANCE TESTING. Because statistical significance is so heavily influenced by SAMPLE size, it is possible for a very weak and, perhaps, trivial effect to be statistically significant thanks to a large sample. On the other hand, a strong effect might go undetected by a test of statistical significance if the sample is very small. Thus, in addition to using tests of statistical significance, one does well to base one's conclusions on effect size, which is designed to be independent of sample size.

The fact that effect sizes are presented in standardized, rather than raw score, units also is important, especially in the social sciences. Many of our measures have arbitrary scales of measurement (e.g., LIKERT SCALES). Because the scales are arbitrary, they have no inherent meaning and may vary across studies. Suppose that the job satisfaction scores of employees after a given workplace intervention were 0.5 units higher, on average, than before the intervention. The recipient of this information can make no sense out of it, if for no other reason than the fact that job satisfaction could be measured on an infinite number of scales. The value of 0.5 might be an enormous change or nearly no change at all. If, on the other hand, the results were reported in a form such as "Postintervention job satisfaction was 0.5 STANDARD DEVIATIONS higher than it was before the intervention," it would be much easier to draw conclusions. Whatever the standard deviation (i.e., whatever the scale), satisfaction was half of a standard deviation higher postintervention. A study attempting to replicate this finding might use a measure of job satisfaction that uses a different scale, but if the results are reported in standard deviation units, then they will be directly comparable to those of any other such study. It is for this reason that effect sizes usually are standardized.

Although there exist many different types of effect size index, a few in particular are worth mentioning here. For correlational data, the most common effect size is the square of the CORRELATION coefficient. The coefficient itself might be a simple correlation, a multiple correlation, or a partial correlation, among other types. The square of a correlation is particularly handy because it reflects the percentage of VARIANCE in one VARIABLE accounted for by another.

For procedures for finding out the DIFFERENCE OF MEANS, such as those associated with ANALYSIS OF VARIANCE designs, there are various available indices. The most common is Cohen's d statistic,

which is defined in a variety of ways, the most common being the difference between two means divided by the pooled standard deviation, s_{pooled} , where s_{pooled} is a function of the within-group variances and sample sizes.

$$s_{\text{pooled}} = \left\{ \frac{[(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2]}{(N_1 + N_2 - 2)} \right\}^{.5}.$$

The resulting d value is simply the mean difference if the dependent variable were scaled to have unit variance. Another way of interpreting d is that it is the proportion of scores in one group that are less than the average score of the other group.

It should be noted that both of the examples of effect size indices given here reflect magnitude of effect and are standardized.

—Jose M. Cortina

EFFECTS CODING

Effects coding is a technique for coding $j - 1$ DUMMY VARIABLES to fully represent the information contained in a j -category classification. Coefficients for effects-coded dummy variables allow the researcher to compare predicted values for designated categories to an "average" value for the overall sample. This "average value" is the unweighted mean of group means, calculated as $[\sum_{j=1}^J \bar{X}_j] \div J$. Only when subgroup sizes are equal does the unweighted mean of group means equal the grand mean (Hardy, 1993). Regression coefficients therefore quantify the difference between the mean of the specified subgroup and the unweighted mean of all subgroup means.

To illustrate, consider the relationship between marital status (married, separated, divorced, widowed, never married) and hourly wages (in US\$). In this example, marital status consists of five (J) subgroups, represented by four ($J - 1$) dummy variables. Effects coding for the dummy variables is illustrated in Table 1. We select married as the omitted category (see Hardy, 1993), which is coded -1 on all four dummy variables. Each dummy variable codes observations in one of the remaining four subgroups "1" and observations in the other three subgroups "0." For example, EFFNM (the effects-coded dummy variable indicating those who never married) codes the never married "1," those currently married " -1 ," and remaining groups (separated, widowed, and divorced) 0.

Table 1

Subgroups	Example of Effects-Coded Dummy Variables				Means (US\$)	N of cases
	EFFNM	EFFS	EFFW	EFFD		
Married	−1	−1	−1	−1	15.33	1,661
Never married	1	0	0	0	12.95	613
Separated	0	1	0	0	11.64	146
Widowed	0	0	1	0	12.94	390
Divorced	0	0	0	1	11.93	14
All cases					14.26	2,773

Table 2

Model	B	Std. Error	t	Sig.
(Constant)	12.96	0.471	27.527	0.000
EFFNM	−0.005	0.536	−0.010	0.992
EFFS	−1.318	0.704	−1.871	0.061
EFFD	−0.022	0.569	−0.038	0.970
EFFW	−1.027	1.755	−0.585	0.559

Consider descriptive statistics and *zero-order correlations* for effects-coded dummy variables. The mean for each effects-coded dummy variable reports the discrepancy in subgroup size between the reference group (coded −1) and the subgroup coded 1, as in $(n_j - n_{\text{ref}})/N$. A negative sign indicates that the reference group has more observations than subgroup j , whereas a positive mean indicates that subgroup j has more observations than the reference group, and the magnitude of the mean indicates the discrepancy in the group sizes relative to the total sample size, or the proportion of excess cases.

When effects-coded dummy variables are used as independent variables in ordinary least squares regression, the *constant* reports the unweighted mean of the subgroups' means. The *coefficients* report each subgroup mean's deviation from this reference point. Sample regression results are reported in Table 2 for a regression of hourly wages on effects-coded dummy variables for marital status. Regression coefficients in Table 2 are unstandardized.

The grand mean for the full set of observations ($N = 2,773$) is 14.26. Because subgroups are unequal in their numbers of observations, however, the unweighted mean of the subgroup means is 12.96, found by adding the five subgroup means and dividing by 5. This value is reported by the constant. When we evaluate the t -tests, we see that the subgroup means for those widowed, divorced, separated, and never married do

not significantly differ from this value. For example, separated's (EFFS) coefficient of −1.318 indicates that the mean hourly wage for separated people is \$1.32 less than the unweighted mean of \$12.96, or \$11.64. Although a difference of more than \$1.00 per hour may seem sizable, the standard error indicates that this wage gap is not reliably measured (e.g., the 95% confidence interval is −1.70 to 1.60). However, the F test indicates that the null hypothesis (which asserts that hourly wage does not differ by marital status) can be rejected at better than the 0.001 alpha level because those currently married have an average hourly wage of \$15.33 $[-1(-0.005 - 1.318 - 0.022 - 1.027) + 12.960 = 15.33]$, which is significantly different from \$12.96 as well as significantly different from the average hourly wage of all other subgroups.

—Melissa A. Hardy and Chardie L. Baird

See also CODING

REFERENCES

- Cohen, J., & Cohen, P. (1983). *Applied multiple regression* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
 Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage.

EFFECTS COEFFICIENT

The total effect, direct and indirect, of a variable in a path model is termed the effects coefficient. Suppose the following RECURSIVE PATH model:

$$Y = a + bX + e,$$

$$Z = c + dX + fY + u.$$

The variable X has a direct effect on Z , estimated by d , and an indirect effect on Z via the influence on Y , estimated by $b \times f$, or bf . The effects coefficient of X

on Z therefore is $d + bf$. Effects coefficients enable the analyst to take into account indirect as well as direct effects in a CAUSAL system, arriving at a summary of the impact of a variable as it works its way through a causal system. However, if that variable is not truly exogenous, calculation of its effects coefficient can be problematic.

—Michael S. Lewis-Beck

See also PATH ANALYSIS

EFFICIENCY

An efficient estimator has a PROBABILITY distribution with the smallest VARIANCE. For example, W and Z are ESTIMATORS of the parameter and their variances are $\text{var}(W)$ and $\text{var}(Z)$, respectively. If $\text{var}(W) < \text{var}(Z)$, then W is a more efficient estimator of θ than is Z . Often, this is referred to as relative efficiency because efficiency requires making a comparison between two or more estimators. Efficiency is a desirable property for estimators because the estimator with smaller variance has a greater chance on any given try of producing the true value θ , given that the estimator is UNBIASED.

To visualize efficiency, think of two people playing darts. The goal is to hit as many bull's eyes as possible. Each person represents a different estimator, and the bull's eye is the mean value of the probability distribution for each estimator. Player 1 throws 100 darts that all fall within a 5-inch circle of the bull's eye. Player 2 also throws 100 darts, but they land within a 12-inch circle of the bull's eye. Player 1 is more efficient than Player 2 because the darts of Player 1 lie closer to the mean (bull's eye). Player 1 also has a higher probability of hitting the bull's eye than does Player 2 because of the smaller variance in the throwing ability of Player 1, given that both players' throws are unbiased (that is, they are as likely to be on one side of the bull's eye as the other, and by the same distance).

An efficient estimator will not always produce an estimate equal to the true value of θ because efficiency is a property of the probability distribution from which the estimates are drawn. This means that in repeated SAMPLING even an efficient estimator will fail to produce an estimate equal to the true θ . However, by using an efficient estimator, one has a higher level of certainty that the estimate is close to the true value for θ . Additionally, an efficient estimator will fail to produce an

estimate of θ equal to its true value if the estimator is biased.

There are so many different estimators that tend to put restrictions on the categories of estimates that we compare to determine their relative efficiency (Kennedy, 1993). First, we look at only unbiased estimates. Unbiased estimators on average produce estimates of θ equal to the true value of the parameter θ . By limiting our comparisons to only unbiased estimators, we compare distributions with the same mean and that produce parameter estimates on average equal to the true value of the parameter. Many times, it is difficult to determine which estimator is the most efficient among several unbiased estimators. An additional restriction can be added, limiting comparisons to only linear estimators. Linear, unbiased estimators with the smallest variance (efficient) represent a special class of estimators called BEST LINEAR UNBIASED ESTIMATOR (BLUE).

—Heather L. Ondercin

REFERENCES

- Kennedy, P. (1993). *A guide to econometrics* (3rd ed.). Cambridge, MA: MIT Press.

EIGENVALUES

Eigenvalues, also known as characteristic roots or latent roots, are a special set of scalars related to MATRIX equations and are an important mathematical concept in a number of statistical methods including FACTOR ANALYSIS and PRINCIPAL COMPONENTS ANALYSIS.

Let $\mathbf{A}$ be a $k \times k$ square matrix of complex numbers; a scalar λ that belongs to the set of complex numbers is said to be an eigenvalue of $\mathbf{A}$ if there exists a nonzero $k \times 1$ column vector $\mathbf{X}$ such that

$$\mathbf{A}\mathbf{X} = \lambda\mathbf{X}.$$

This column vector $\mathbf{X}$ is known as the eigenvector of $\mathbf{A}$. Because it is positioned to the right of $\mathbf{A}$, it is called a "right eigenvector." A row eigenvector that is positioned to the left of $\mathbf{A}$ is called a "left eigenvector." Each eigenvalue is associated with a pair of eigenvectors—a left and a right eigenvector. The decomposition of $\mathbf{A}$ into eigenvalues and eigenvectors is known as eigen decomposition. The equation above also can be expressed more compactly as

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{X} = 0,$$

where $\mathbf{I}$ is the identity matrix. When an eigenvalue is distinct from all other eigenvalues, its eigenvector is unique. As an example of the application of eigenvalues, an eigenvalue in a dimension from a principal components analysis measures the goodness of fit and gives the proportion of variance in the original variables accounted for by the principal component.

—Tim Futing Liao

See also FACTOR ANALYSIS

EIGENVECTOR. See EIGENVALUES;
FACTOR ANALYSIS

ELABORATION (LAZARSFELD'S METHOD OF)

In its most elementary form, application of the elaboration method implies that one starts with a two-dimensional table showing the relationship between two variables and then one introduces a third variable for elaboration and elucidation. Table 1 [inspired by the basic text by Lazarsfeld (1955) and by Hyman (1955)] serves as an example dealing with the (fictitious) relationship between the dichotomous variables X -Sex (1 = Woman, 2 = Man) and Y -Car accident last year (1 = Yes, 2 = No). The table is presented in percentages, with the absolute numbers between parentheses.

Table 1

<i>Y</i> -Car Accident	<i>X</i> -Sex		Total
	1 = Man	2 = Woman	
1 = Yes	38% (100)	19% (50)	28% (150)
2 = No	62% (160)	81% (220)	72% (380)
Total	100% (260)	100% (270)	100% (530)

It appears that women are better drivers than men: Whereas 38% of the men were involved in a car accident last year, this is true for only 19% of the women. What is the cause of this difference? Are women more careful, considerate drivers? A simple explanation might be that men, in the past, drove many more miles

than women. It stands to reason that the more a person drives, the greater that person's chances are of getting involved in a road accident. The logic underlying the elaboration method is that if this latter interpretation were true, comparing men and women who drive the same number of miles per year would result in no difference in the number of accidents between men and women. We introduce a third variable, T -Mileage (1 = High, 2 = Low) and investigate the relationship between X and Y within the categories of the third variable (we "control for T " or "we hold T constant"). Introduction of the third variable is shown in Table 2.

Within the subtable for high mileage ($T = 1$), people have many more accidents than in the subtable for low mileage ($T = 2$), but there is no longer a difference between men and women within these subtables (i.e., when comparing men and women who drive the same amount of miles). Lazarsfeld called this type of outcome "interpretation." Representing direct causal effects by an arrow, the causal scheme of this outcome is $X \rightarrow T \rightarrow Y$, meaning X is a cause of Y but indirectly through T (as a result of all kinds of social circumstances, women drive less than men and *therefore* are involved in fewer accidents). If there were other differences left between men and women that caused women to be less involved in accidents than men, then even in the subtables controlling for T , there would have been differences in percentages between men and women.

Another type of outcome is SPURIOUSNESS. For this type, the original ASSOCIATION also disappears, as above, but now the causal order of X and T is changed: X is seen as a consequence of T rather than as a cause. The causal scheme is $X \leftarrow T \rightarrow Y$. Let us assume that for a table with X -Education and Y -Conservatism we find that less-educated people are more conservative than people with more education. At the same time, it may be true that (for T -Age) the older one gets, the more conservative one becomes; furthermore, younger people may on average be better educated. If age is responsible in this way for the original relationship between education and conservatism, the original relationship will disappear when controlling for age in the manner described above. From the interpretation outcome we learn that X and Y are indirectly causally related to each other; spuriousness unmasks an original CORRELATION and shows that the relationship is noncausal: Age (T) causes X (Education), and Age also causes Conservatism (Y), but X in no way causes Y .

Table 2

Y-Car accident	T-Mileage (1 = High)			T-Mileage (2 = Low)		
	X-Sex		Total	X-Sex		Total
	1 = Man	2 = Woman		1 = Man	2 = Woman	
1 = Yes	50%	50%	50%	7%	8%	7%
	(95)	(35)	(130)	(5)	(15)	(20)
2 = No	50%	50%	50%	93%	92%	93%
	(95)	(35)	(130)	(65)	(185)	(270)
Total	100%	100%	100%	100%	100%	100%
	(190)	(70)	(260)	(70)	(200)	(270)

The third type of elaboration is specification or INTERACTION. In this type of elaboration, the relationships between X and Y are different in the subtables. For example, in the first subtable above (for high mileage), it might occur that men cause more accidents than women, but when both drive less (low mileage), there may be no difference between men and women. Interaction specifies how the (causal) relationship between two variables differs within the categories of a third variable.

The elaboration method fulfills, for categorical variables, much the same role as MULTIPLE REGRESSION ANALYSIS fulfills for CONTINUOUS VARIABLES. Its basic principles can be extended to more than three variables and to more complicated situations, although a more formal approach is then required. Such a formal approach has been developed within the context of the LOG-LINEAR MODEL. Lazarsfeld's ultimate aim was to develop a sound approach toward CAUSAL MODELING using "sociological" data, meaning (for him) CATEGORICAL data obtained by means of non-experimental survey designs. Many of his insights into the problems and possibilities of causal analysis can be found in modern treatments of causality, such as Pearl (2000).

—Jacques A. Hagenaars

REFERENCES

- Hagenaars, J. A. (1998). Categorical causal modeling: Latent class analysis and directed log-linear models with latent variables. *Sociological Methods and Research*, 26, 436–486.
- Hyman, H. (1955). *Survey design and analysis*. New York: Free Press.
- Lazarsfeld, P. F. (1955). Interpretation of statistical relations as a research operation. In P. F. Lazarsfeld & M. Rosenberg (Eds.), *The language of social research* (pp. 111–125). Glencoe, IL: Free Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

ELASTICITY

Social science analysts often are interested in knowing the effect that a change in one variable produces on another. One useful way of capturing the direction and magnitude of this effect is the concept of *elasticity*, which is defined as follows: The elasticity of a function is the proportionate change in the dependent variable divided by the proportionate change in the independent variable at given values of both variables. That is,

$$E = \frac{\% \text{ change in } y}{\% \text{ change in } x} = \frac{(\Delta y/y) \times 100}{(\Delta x/x) \times 100} = \frac{\Delta y}{\Delta x} \times \frac{x}{y}, \quad (1)$$

where the symbol Δ before a variable means a discrete change in the value taken by the variable and y and x are the DEPENDENT VARIABLE and an INDEPENDENT VARIABLE, respectively.

As an example, suppose that we are interested in measuring the effect of the consumption of calories on a person's weight under a particular diet. Let us assume that y is a person's weight in kilograms and x is the consumption of calories measured in kilocalories per day. Then, if we use absolute changes, the answer to the previous question would be arbitrarily affected by the choice of units. That is, if the per-day consumption of calories increases from 1,000 kilocalories to 1,100 kilocalories and people, as a result, increase their weight from 60 to 63 kilograms, we get the impression that the effect is quite small. After all, an increase of 100 kilocalories produces a

weight gain of only 3 kilograms. However, by changing the unit of measure for weight from kilograms to grams, we find that an increase of 100 kilocalories causes a weight gain of 3,000 grams. We now get the impression that the effect is large. Using the ratio between percentage changes avoids this problem. The given weight gain is 5% whether measured in terms of kilograms $[(3\text{k}/60\text{k}) \times 100]$ or in terms of grams $[(3,000\text{k}/60,000\text{k}) \times 100]$.

It is important to note that if the variation in the value taken by the variables is not small, it is common to use an average between the values taken by the variables before and after the variation is produced. In this case, the elasticity takes the form of the following expression:

$$E = \frac{\Delta y}{\Delta x} \times \frac{\frac{(x_1 + x_2)}{2}}{\frac{(y_1 + y_2)}{2}} = \frac{\Delta y}{\Delta x} \times \frac{x_1 + x_2}{y_1 + y_2}. \quad (2)$$

The subscripts 1 and 2 indicate the values taken by the variable before and after the variation is produced, respectively. Notice that if the variation in the value taken by the variables is not small, we will have the problem that the elasticity in equation (1) will be different under an increase in the consumption of calories than under a decrease. That is, when the consumption of calories rises from 1,000 to 1,100 kilocalories, the elasticity would be

$$\frac{3}{100} \times \frac{1,000}{60} = 0.5.$$

When the consumption of calories falls from 1,100 to 1,000 kilocalories, it would be

$$\frac{3}{100} \times \frac{1,100}{63} = 0.524.$$

By using expression (2), however, we get an estimate of the elasticity equal to 0.512 whether there is an increase or a decrease.

Furthermore, when Δ is very small (i.e., an infinitesimal increase), the previous formulas converge to one in which derivatives are used. That is,

$$E = \frac{\partial y}{\partial x} \times \frac{x}{y}. \quad (3)$$

The concept of elasticity is also very important in REGRESSION analysis. For instance, when the regression is a linear one, like $y = \alpha + \beta x + \varepsilon$, the use

of equation (3) gives an elasticity equal to $\beta(y/x)$. However, in the case that the regression is a logarithmic or a semilogarithmic one, we have the following cases:

$$\ln y = \alpha + \beta_1 \ln x_1 + \beta_2 x_2 + \varepsilon$$

$$\text{use } \begin{cases} E_{y,x_1} = \frac{\partial \ln y}{\partial \ln x} = \beta_1 & (4a) \\ E_{y,x_2} = \frac{\partial \ln y}{\partial x} \cdot x = \beta_2 x_2 & (4b) \end{cases}$$

and

$$y = \alpha + \beta \ln x + \varepsilon \quad \text{use} \quad E_{y,x} = \frac{\partial y}{\partial \ln x} \cdot \frac{1}{y} = \beta \frac{1}{y}. \quad (5)$$

Let's try to see how it works by using an example taken from Lindahl's (2002, p. 7) study of intergenerational income mobility in Sweden between 1962 and 1973. The purpose of Lindahl's study was to obtain a measure of the degree of equality of opportunities, as represented by intergenerational income mobility. Here we interpret only the case of families with only one male son. The regression results are as follow:

$$y = 8.29660 + 0.227x_1 + 0.05712x_2 - 0.00067955x_3, \quad (6)$$

where y is the long-run log income of the sons, x_1 is long-run log income of the fathers, x_2 is the fathers' age, and x_3 is the square of the fathers' age. The reason for the inclusion of the fathers' age in the regression is to correct for the fact that the long-run income of a person probably is higher at the age of 50 than at the age of 25. Conversely, it was not necessary to correct for the sons' age because every son in the study was of the same age.

We can apply formula (4a) in the previous example to estimate the intergenerational income elasticity, which is equal to 0.227. This elasticity implies that, controlling for fathers' age effects, a 1% rise in the long-run income of the fathers produces a 0.227% rise in the long-run income of the sons.

For further readings on elasticity, see Eatwell et al. (1987).

—Osiris Jorge Parcerro

REFERENCES

Eatwell, J., Milgate, M., & Newman, P. (1987). *The new Palgrave: A dictionary of economics*. London: Macmillan.

Lindahl, L. (2002). *Do birth order and family size matter for intergenerational income mobility? Evidence from Sweden* (Working Paper No. 5). Stockholm: Swedish Institute for Social Research.

EMIC/ETIC DISTINCTION

Linguist Kenneth L. Pike, in 1954, coined the terms *emic* and *etic* from *phonemic* and *phonetic*. Pike used *emic* to refer to the intrinsic cultural distinctions meaningful to the members of a cultural group and *etic* to refer to the extrinsic ideas and categories meaningful for researchers.

For example, modern medical science defines “diseases” in precise, culture-free (*etic*) ways for any patient, whereas traditional peoples define “illnesses” differently based on their particular cultural contexts (*emically*). Western botanists categorize flora and fauna *etically*, based on the Linnaean taxonomy, whereas indigenous peoples categorize them *emically*, based on their particular folk-science worldview. Some cultures in Papua, New Guinea (PNG) classify bats with “birds” rather than “mammals” because they fly like most birds. Optical physicists divide the color spectrum according to the wavelength of light *etically*, whereas local peoples divide it in different ways into *emically* meaningful divisions based on their languages and cultures. For example, the Selepet people of PNG use a single color term that includes the English *yellow*, *orange*, and *brown*.

Anthropologist Marvin Harris borrowed the terms from Pike in 1964 and applied them to his own CULTURAL MATERIALISM method of studying cultures. Misunderstanding Pike, he used the terms very differently. Pike had argued for an *emic* approach to studying a culture if one wanted to understand it correctly. In contrast, Harris defined the *emic* method as referring to the ideal reasons the natives give for their customs, and he used *etics* to refer to the subconscious real reasons for customs. For example, the Yanomami Indians say they raid enemy villages to kidnap young women (an *emic* explanation, in Harris’s definition); but the real *etic* reason, which the Indians don’t recognize, is that the enemy villages are encroaching on the scarce wild game resources needed by the raiding group.

COGNITIVE ANTHROPOLOGISTS follow Pike in using the *emic/etic* concept as a research method, whereas cultural materialists use the method as Harris

remodeled it. This resulted in some confusion among anthropologists and linguists by the 1970s as to what *emics* and *etics* were supposed to mean. By the 1980s, the terms were appearing in many of the writings of social scientists. The meanings of the terms had multiplied in many more directions, in part because the *emic/etic* concept was now being used in other academic disciplines (including psychology, sociology, medicine, psychiatry, economics, and religion). These new meanings are described by Headland, Pike, and Harris (1990, pp. 15–23).

One commonality in both Pike’s and Harris’s models of *emics* and *etics* was that they both agreed that an *emic* description of a culture is not necessarily a model of which the natives are consciously aware. Even though many definitions refer to the *emic* model as the insider’s view, that is not always so. Whereas *emic* constructs may make sense to those who hold them, those people often are not aware of them unless an outside ethnographer explains their own *emic* constructs to them.

—Thomas N. Headland and
Kenneth A. McElhanon

See also COMPONENTIAL ANALYSIS

REFERENCES

- Harris, M. (1976). History and significance of the *emic/etic* distinction. *Annual Review of Anthropology*, 5, 329–350.
- Headland, T. N., Pike, K. L., & Harris, M. (1990). *Emics and etics: The insider/outsider debate*. Newbury Park, CA: Sage.
- Jahoda, G. (1977). In pursuit of the *emic-etic* distinction: Can we ever capture it? In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural research* (pp. 55–63). Amsterdam: Swets & Zeitlinger.
- Lett, J. (1987). The importance of the *emic/etic* distinction. In *The human enterprise: A critical introduction to anthropological theory*. Boulder, CO: Westview.
- Pike, K. L. (1967). *Language in relation to a unified theory of the structure of human behavior* (2nd ed.). The Hague: Mouton. (Original work published 1954.)

EMOTIONS RESEARCH

Emotions research focuses on understanding the social, psychological, and physiological processes involved in emotional experience; the social construction of emotions (including historically and cross-culturally); the role of emotions in social life; the

management of emotions in social interaction; and the role of emotions in the research process itself.

Researchers in this subfield collect data on emotions via methods that may closely resemble those used in other branches of sociology and social psychology, among them SURVEYS, EXPERIMENTS, FIELD RESEARCH, and IN-DEPTH INTERVIEWS. Their analyses entertain a variety of qualitative and quantitative techniques. A few approaches bear special mention.

In research based on affect control theory (ACT) (Heise, 1987), emotions are defined as arising from the actions and the confirmed or disconfirmed identities of people involved in interactions. Social actors may respond to the emotional cues that result from disconfirmed identities by redefining situations or other people's identities. Data collection for ACT frequently involves asking participants to rate concepts on three dimensions—evaluation (good vs. bad), potency (powerful vs. weak), and activity (lively vs. inactive)—using a SEMANTIC DIFFERENTIAL SCALE. The majority of analyses using ACT are quantitative.

In qualitative, field- and interview-based studies of emotions, researchers note social actors' emotional displays as they occur in context and record participants' descriptions of felt emotions that could not be inferred from observation alone. Researchers also note the consequences that unfold from emotion-related interactions. Those who apply CONVERSATION ANALYSIS techniques to study emotions in face-to-face communication may examine highly specific exchanges. Field researchers and in-depth interviewers, by contrast, explore a range of situations and conditions in order to understand how emotions work in conjunction with other social concepts and processes such as identity, perception, social relationships, social status, group dynamics, or social movements.

Arising from a critical reaction to the "affective neutrality" that social scientists aspired to maintain in the past, many field researchers now include themselves as research participants rather than acting solely as detached, unemotional observers. They treat their emotional experiences during fieldwork as valuable data, and they incorporate reflexive observations and insights in their FIELDNOTES and qualitative analyses (Kleinman & Copp, 1993). Researchers who include emotions as data can gain deeper sociological insights about the participants and the settings they study, the social constraints under which participants operate, and the social implications of their research.

To study the lived experience of emotions and to better understand how people experience emotions as complex processes, field researchers may employ the sociological technique of *systematic introspection* (Ellis, 1991). This method involves recording one's emotional reactions to people, objects, and events and then analyzing what those feelings mean and how they inform ongoing social situations and cultural practices. Systematic introspection may be conducted by trained participants or practiced by researchers as an AUTOETHNOGRAPHIC research tool.

—Martha Copp

See also REFLEXIVITY

REFERENCES

- Ellis, C. S. (1991). Sociological introspection and emotional experience. *Symbolic Interaction*, 14, 23–50.
- Heise, D. R. (1987). Affect control theory: Concepts and model. *Journal of Mathematical Sociology*, 13, 1–33.
- Kleinman, S., & Copp, M. A. (1993). *Emotions and fieldwork*. Thousand Oaks, CA: Sage.

EMPIRICISM

Like "POSITIVISM," the term "empiricism" has come to be used by social scientists in ways that often indicate little more than negative evaluation. Indeed, these two words are frequently treated as synonyms. Both are now closely associated with the discredited idea that science is the only genuine source of knowledge because its conclusions are logically derived from empirical data.

In EPISTEMOLOGY, "empiricism" refers to a distinctive approach which claims that all knowledge comes from the senses, by contrast with the RATIONALIST argument that there is innate knowledge (see Hamlyn, 1967). In large part, empiricism arose out of the scientific revolution of the 17th century, with its emphasis on empirical (especially experimental) investigation of the physical world as the key to scientific knowledge. This was a departure from earlier ideas about the process of inquiry, in which reliance on ancient texts, religious faith, and/or systematic philosophizing had been influential.

It is perhaps worth emphasizing the radical and progressive character of 17th-century empiricism. This arose from the fact that it involved a critique both of

previous claims to knowledge and of the means by which these had been justified. Moreover, it demanded that any criticism be answered in terms of an appeal to evidence of a kind that was, in principle, available to anyone. In this way, all forms of authority were opened up to potential challenge.

The meaning and validity of empiricist theories of knowledge have been matters of recurrent philosophical dispute. Against them, it has been argued that no data are simply available to researchers or observers without judgment or interpretation, and there are, therefore, no data whose validity can be taken as indubitable. Equally, it is claimed that science does not, and could not, operate simply by seeking logically to infer laws from observational data. Instead, it requires development of theories, which are then tested against data; a process that can only indicate likely truth or falsity, rather than establishing VALIDITY beyond all doubt.

In the context of the social sciences, *empiricism* generally has been used to refer to approaches that are felt to place too much emphasis on empirical data, discouraging theory construction. An example is C. Wright Mills's critique of "abstracted empiricism" (Mills, 1959). Empiricism also has been seen as involving an excessively narrow conception of evidence, one that limits it to what is observable or quantifiable. As this suggests, "empiricist" is a term of criticism most commonly applied to QUANTITATIVE RESEARCH, though there are examples of its application to QUALITATIVE RESEARCH. For example, some kinds of ETHNOGRAPHIC research have been challenged, notably by CRITICAL THEORISTS, for being preoccupied with appearances, that is, with documenting commonsense interpretations of the world, thereby neglecting the underlying structures that generate those ideological appearances. On the other side, although few would seek to defend empiricism in any narrow form, it can be argued that criticism of it has become exaggerated today, encouraging the presentation of speculative theorizing as fact.

—Martyn Hammersley

REFERENCES

- Bryant, C. G. A. (1985). *Positivism in social theory and research*. London: Macmillan.
- Hamlyn, D. W. (1967). Empiricism. In P. Edwards (Ed.), *The encyclopedia of philosophy* (pp. 499–505). New York: Macmillan.

Mills, C. W. (1959). *The sociological imagination*. New York: Oxford University Press.

EMPTY CELL

An empty cell occurs when there are no cases in the CELL of a CONTINGENCY TABLE. Empty cells create problems of interpretation, as do cells containing very small numbers of cases.

—Alan Bryman

See also SPARSE TABLES

ENCODING/DECODING MODEL

The encoding/decoding model of media research, originally formulated by Stuart Hall (1974) at the University of Birmingham's Centre for Contemporary Cultural Studies in the 1970s, has been influential in cultural studies and the sociology of mass communications. It recalls the tripartite structure of the most elementary of communications models—sender, message, and receiver—but emphasizes the relatively autonomous moments of encoding and decoding and the textual polysemy (multiple meanings) of television programs in particular. The meaning of any message is not determined, in the end, by the intention of the communicator. It is not to be assumed, however, that audiences simply misunderstand intended meaning. As Umberto Eco (1965/1972) argued, "aberrant decoding" of mass communications is normal because modern media address such large and socially heterogeneous audiences that inevitably will bring their different frameworks of understanding to bear on the interpretation of messages.

Aberrant decoding of media messages is not just a matter of individual differences, according to Hall, but is also socially motivated. In a capitalist society, class differences are especially important and are represented by contrasting value systems and discourses. Hall appropriated Frank Parkin's (1973) distinction between dominant, subordinate, and radical meaning/value systems in order to examine what he termed dominant, negotiated, and oppositional decodings. It is reasonable to suppose that the mass media typically will present the legitimacy of dominant—in effect, "bourgeois"—values and ways of making

sense of the world. Accordingly, preferred meanings and readings are routinely encoded into media texts, usually representing the ideological interests of the powerful. The relatively powerless may either accept or refuse the dominant ideological framework encoded in, say, television news programs. For instance, regarding the coverage of industrial conflict from the point of view of management, a worker of left-wing persuasion may interpret it as biased. In between taking the preferred reading straight (dominant decoding) or questioning it fundamentally (oppositional decoding), various kinds of negotiated decoding might be made more commonly by audience members.

Hall's encoding/decoding model was a key feature of the paradigm of hegemony theory in cultural and media studies. This was a significant revision of the classical Marxist dominant ideology thesis. Hall wanted to retain, to some extent, the argument that the dominant ideas in society were those of the ruling class while investigating the sheer SEMIOTIC complexity of ideological process through channels of communication. It was evidently the case that subordinate groups in society are not just indoctrinated by the dominant ideology but also are likely to question its validity according to their own interpretive frameworks, experiences, and interests. Ideological power, then, was a matter of discursive negotiation in the struggle for social leadership rather than simply imposed from above.

The encoding/decoding model was a heuristic device for putting hegemony theory into operation in critical media research. A student of Hall, David Morley (1980), applied it to study how, in educational settings, groups from different class backgrounds interpreted a newsmagazine program, *Nationwide*, on British television during the late 1970s. He reached two important conclusions that were influential in the further development of research on encoding and decoding. First, class was neither the only nor necessarily the most important factor in differential decoding. Ethnicity, gender, and generation were key sociological determinants as well. In subsequent work by Morley and his followers, recognition of finely differentiated decoding of media genres proliferated to such an extent that the very notion of a preferred meaning and the ideological critique implicit in it was undermined. Second, Morley (1986) himself became dissatisfied with the artificiality of FOCUS GROUP research, which he had employed in his *nationwide* study, and more interested in the ETHNOGRAPHY of media use, especially gender

relations surrounding television in the home and the idea of the active audience. As the encoding/decoding model developed, then, it became less concerned with questioning the ideological power of mass media and more focused on domesticated micropolitics and, in effect, the vicissitudes of sovereign consumption (see McGuigan, 1992, for a fuller critical assessment).

—Jim McGuigan

REFERENCES

- Eco, U. (1972). Towards a semiotic inquiry into the television message. In *Working papers in cultural studies* (Vol. 3, pp. 103-121). Birmingham: Centre for Contemporary Cultural Studies. (Original work published 1965)
- Hall, S. (1974). The television discourse—Encoding and decoding. In *Education and culture* (Vol. 35, pp. 8-14). Paris: United Nations Educational, Social, and Cultural Organization.
- McGuigan, J. (1992). *Cultural populism*. London: Routledge.
- Morley, D. (1980). *The "Nationwide" audience*. London: British Film Institute.
- Morley, D. (1986). *Family television—Cultural power and domestic leisure*. London: British Film Institute.
- Parkin, F. (1973). *Class inequality and political order*. London: Paladin.

ENDOGENOUS VARIABLE

An endogenous variable is a factor in a CAUSAL MODEL or a causal system whose value is determined by the states of other variables in the system, in contrast with an *exogenous variable*, the value of which is determined outside the system. Related but non-equivalent distinctions are those between DEPENDENT and INDEPENDENT VARIABLES and between explanandum and explanans. A factor can be classified as endogenous or exogenous only relative to a SPECIFICATION of a MODEL representing the causal relationships producing the outcome y among a set of causal factors $\mathbf{X}(x_1, x_2, \dots, x_k)$ where $y = \mathbf{M}(\mathbf{X})$. A variable x_j is said to be endogenous within the causal model $\mathbf{M}$ if its value is determined or influenced by one or more of the independent variables $\mathbf{X}$ (excluding itself). A purely endogenous variable is a factor that is entirely determined by the states of other variables in the system. (If a factor is purely endogenous, then in theory we could replace the occurrence of this factor with the functional form representing the composition of x_j

as a function of $\mathbf{X}_t$.) In real causal systems, however, there can be a range of endogeneity. Some factors are causally influenced by factors within the system but also by factors not included in the model. A given factor therefore may be partially endogenous and partially exogenous—partially but not wholly determined by the values of other variables in the model.

Consider a simple causal system—farming. The outcome we are interested in explaining (the dependent variable or the explanandum) is crop output. Many factors (independent variables, explanans) influence crop output: labor, farmer skill, availability of seed varieties, availability of credit, climate, weather, soil quality and type, irrigation, pests, temperature, pesticides and fertilizers, animal practices, and availability of traction, among others. These variables are all causally relevant to crop yield, in a specifiable sense: If we alter the levels of these variables over a series of tests, the level of crop yield will vary as well (up or down). These factors have real causal influence on crop yield, and it is a reasonable scientific problem to attempt to assess the nature and weight of the various factors. We can also notice, however, that there are causal relations among some of but not all these factors. For example, the level of pest infestation is influenced by rainfall and fertilizer (positively) and pesticide, labor, and skill (negatively). Pest infestation therefore is partially endogenous within this system and partially exogenous, in that it is also influenced by factors that are external to this system (e.g., average temperature, presence of pest vectors, decline of predators).

The concept of endogeneity is particularly relevant in the context of TIME SERIES analysis of causal processes. It is common for some factors within a causal system to be dependent for their value in period n on the values of other factors in the causal system in period $n - 1$. Suppose that the level of pest infestation is independent of all other factors within a given period but is influenced by the level of rainfall and fertilizer in the preceding period. In this instance, it would be correct to say that infestation is exogenous within the period but endogenous over time.

—Daniel Little

See also EXOGENOUS VARIABLE

REFERENCES

- Hendry, D. F. (1995). *Dynamic econometrics*. Oxford, UK: Oxford University Press.

- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

EPISTEMOLOGY

In philosophy, epistemology refers to a THEORY of knowledge, a theory of how human beings come to have knowledge of the world around them—of how we know what we know. Epistemology provides a philosophical grounding for establishing what kinds of knowledge are possible and for deciding how knowledge can be judged as being both adequate and legitimate. In the social sciences, the term is used in the context of deciding which scientific procedures produce reliable social scientific knowledge.

Two theories of knowledge have predominated in philosophical discourse since the scientific revolution in the 17th century: RATIONALISM (represented by René Descartes, Gottfried Leibniz, and Benedict de Spinoza) and EMPIRICISM (represented by John Locke, George Berkeley, and David Hume). The concern was to find a secure foundation for scientific knowledge and to distinguish this from belief and prejudice. Rationalism is based on the idea that reliable knowledge is derived from the use of “pure” reason, from establishing indisputable axioms and then using formal logic to arrive at conclusions. From this point of view, mathematics produces such knowledge. Empiricism, on the other hand, relies on the use of the human senses to produce reliable knowledge. This means that knowledge of the world can be obtained only through direct sense-experience.

In the context of the social sciences, these philosophical positions can be further elaborated in terms of two dominant epistemological positions and their associated ontological positions (see ONTOLOGY). In the first epistemological position, known as nominalism, the concepts that are used in description and explanation are simply regarded as convenient, collective names that are invented as summaries of the general categories of things that have been observed, such as “social actors” or “social groups.” These collectivities should neither be confused with reality itself nor attributed with the capacity to act; reality is made up of events, and only individuals can act. In the second epistemological position, known as REALISM, scientific concepts are viewed as revealing something about social reality that is not necessarily observable. Such concepts are designed to penetrate beyond observable

events to a reality that underlies and explains them (see also CRITICAL REALISM).

When nominalism and realism are combined with the two major alternative ontological positions, materialism and idealism, a four-way classification scheme is generated. Empiricism combines a materialist ontology (see IDEALISM) with a nominalist epistemology, substantialism combines a materialist ontology with a realist epistemology, subjectivism combines an idealist ontology with a nominalist epistemology, and rationalism combines an idealist ontology with a realist epistemology (Johnson, Dandeker, & Ashworth, 1984).

In empiricism, reality is viewed as being constituted of material things that can be observed by the use of the human senses. Concepts and generalizations are shorthand summaries based on many observations. Substantialism also adopts a materialist view of reality but accepts that people in different times and places can interpret reality differently. Nevertheless, the material world is seen to constrain human actions and social relations. Because subjectivism rejects the notion of a material world and views reality as being socially constructed and interpreted, knowledge of this reality is available only from the accounts that social actors can give of it. Finally, rationalism views reality as both real and general; it exists independently of people, their consciousness, and their circumstances. Because this reality is made up of ideas, knowledge of it can be obtained only by examining thought process, the innate ideas shared by human beings—in short, the structure of mind itself. These four positions must be regarded as ideal types, between which there are inherent tensions (Johnson et al., 1984). They are associated with the major PHILOSOPHIES OF SOCIAL SCIENCE. Empiricism is associated with POSITIVISM and FALSIFICATIONISM, substantialism is associated with CRITICAL REALISM, and subjectivism is associated with INTERPRETIVISM. Rationalism can be found in Emile Durkheim's work on suicide but is now uncommon in the social sciences.

—Norman Blaikie

REFERENCES

- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
 Crotty, M. (1998). *The foundations of social research*. London: Sage.

- Johnson, T., Dandeker, C., & Ashworth, C. (1984). *The structure of social theory*. London: Macmillan.

EQS

EQS is a statistical software package for performing analysis of STRUCTURAL EQUATION MODELS. It is known for its ease in handling non-normal and non-continuous data. For further information, see the Web site of the software: <http://www.mvsoft.com/>

—Tim Futing Liao

ERROR

In everyday usage, “error” means “mistake.” In arithmetic, “error” means the degree of inaccuracy in a calculation. In statistics, the term “error” may not have a negative connotation. It generally represents the difference or discrepancy between the value of an entity observed (e.g., the sample mean) and its “true” or expected value (e.g., the population mean). Very often, this difference is due to the result of chance elements. That is why it is often called a stochastic error or random error term. (The word “stochastic” comes from the Greek word *stokhos*, meaning the “bull’s eye.” The outcomes of throwing darts onto a dart board is a stochastic process, that is, a process fraught with misses. In statistics, the word implies the presence of a random variable—a variable whose outcome is determined by a chance experiment.)

There are several types of errors that one may encounter in practice. These include WHITE NOISE error, TYPE I ERROR, TYPE II ERROR, errors in MEASUREMENT, errors in variables, errors in equations, model specification error, errors resulting from missing observations, and FORECASTING errors.

WHITE NOISE ERROR

Let u_t be a random error term. If it has a zero mean, constant variance, and no autocorrelation (or SERIAL CORRELATION), it is called a white noise error term, denoted by $u_t \sim WN(0, \sigma)^2$. If in addition to being serially uncorrelated, u_t is serially independent, then u_t is independent white noise, denoted by $u_t \sim iid(0, \sigma)^2$, where *iid* means independently and

identically distributed. If, in addition, u_t is normally distributed, then it is said to be normal white noise or Gaussian white noise, denoted by $u_t \sim iidN(0, \sigma)^2$. It may be noted that zero correlation means independence only in the normal case.

The white noise error term plays a crucial role in TIME SERIES modeling (see FORECASTING) as well as in the classical linear regression model.

ERRORS OF OBSERVATION OR MEASUREMENT

In collecting data, errors often arise because of faulty measurement instruments or because of human factors. A common error is *rounding error*, or *error of approximation*. In this type of error, instead of recording observations, say, to the fourth decimal point, one rounds them to three decimal points. In survey-type data, the questioner may not obtain information on all the questions because of nonresponse, or not all questionnaires may be returned. Sometimes information on variables such as income and wealth is difficult to come by because respondents are reluctant to disclose the true information for fear of creating problems with tax authorities.

ERRORS IN VARIABLES

Sometimes information on variables is inherently difficult to obtain. For instance, data on expected, or anticipated, inflation is hard to come by because one person's expectation of inflation may not coincide with another person's. Economists have devised various schemes, or *proxy variables*, to measure the anticipated or expected inflation. One measure of expected inflation is obtained by surveying leading economists about their views on inflation expected in the near future. The views of these economists may be based on mathematical models or may be purely personal. To take another example from economics, the variable "natural rate of unemployment" is very difficult to measure. Some economists take some kind of weighted average of the actual rate of unemployment in the past several months or quarters and use that as a proxy for the natural rate of unemployment. All these proxy measures are likely to contain errors.

ERRORS IN TESTING HYPOTHESES

In testing statistical hypotheses, one may commit a TYPE I ERROR or a TYPE II ERROR. A Type I error consists

of rejecting a true hypothesis, whereas a Type II error consists of not rejecting a hypothesis even though it may be false. Both these types of errors are fundamental to the Neyman-Pearson, or classical, theory of hypothesis testing. Ideally, one would like to minimize both these types of errors; however, for a *given* sample size, it is not possible to minimize both types of errors simultaneously. Classical hypothesis testing usually fixes the probability of a Type I error at some arbitrary LEVEL OF SIGNIFICANCE, such as 1%, 5% or 10%, and does not worry much about a Type II error; In some respects, a Type I error is regarded as more "serious" than a Type II error, in the sense that convicting an innocent person is more serious than letting a guilty person go free.

EQUATION ERROR

Equation error typically is associated with REGRESSION analysis. In regression analysis, the objective is to explain the behavior of one variable, Y (the DEPENDENT VARIABLE), in terms of one or more X variables (the EXPLANATORY VARIABLES). No matter how many X variables are considered, one can never explain the behavior of Y completely in terms of the chosen X variables. The difference between the actual values of Y and those estimated from the regression model represents the equation error term. Symbolically, the regression model may be written as

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i.$$

What this model states is that regardless of the number of explanatory variables (the X s) considered in the model, there is bound to be some variation in Y that cannot be explained by all the X variables. There is bound to be some intrinsic randomness in Y that cannot be explained, and this intrinsic randomness is represented by the STOCHASTIC error term, u_i . This error term may be due to several factors, such as vagueness of the theory underlying the regression model, unavailability of data on some explanatory variables, or use of a poor proxy variable (e.g., using the actual inflation rate instead of the expected inflation rate in models of unemployment). Sometimes, a researcher may know all the explanatory variables in a model but, following the principle of parsimony (or Occam's razor), may decide to include only a few core variables in the analysis.

The stochastic error term u_i plays an extremely critical role in regression analysis because the

properties of the estimated regression coefficients—the β coefficients—depend critically on the assumptions about the stochastic error term. In the classical regression model, it is assumed that the values of the X , or explanatory, variables are fixed in repeated sampling, and the error term u_i is assumed to be Gaussian white noise. As a result, the estimated β coefficients also follow the normal distribution. This makes the task of hypothesis testing about the true, or population, β coefficients much simpler. Also, under the normality assumption about the error term u , the MAXIMUM LIKELIHOOD ESTIMATION (MLE) and ORDINARY LEAST SQUARES (OLS) estimators of the β coefficients are the same.

It should be noted that in practice one does not observe the error term u_i . Instead, what one observes is the RESIDUAL term $\hat{u}_i$, which is the difference between the actual value of Y_i and its value estimated from the regression model. In other words, $\hat{u}_i$ is a proxy for the true u_i . It can be shown that as the sample size increases indefinitely, the residuals tend to converge to their true values.

MODEL SPECIFICATION ERRORS

A theory may suggest the variables that may be included in a model, but the theory often is unable to suggest the specific functional form in which the variables should enter the model. Consider, for instance, the demand for a product (Y) in relation to its price (P) and the income of the consumer (I). Now consider two possible (regression) models:

$$Y_i = \beta_1 + \beta_2 P_i + \beta_3 I_i + u_i,$$

$$\ln Y_i = \alpha_1 + \alpha_2 \ln P_i + \alpha_3 \ln I_i + u_i,$$

where u_i is the stochastic error term and “ln” stands for the natural logarithm.

The first model is a linear regression model, whereas the second is a log-linear regression model. Although both models contain the same variables, the way they enter the model makes a big difference in the interpretation of the model. In the linear model, for instance, β_2 (the partial slope coefficient) measures the rate of change in the mean value of Y with changes in the value of P , whereas in the log-linear model α_2 measures the *elasticity* of Y with respect to P ; that is, it measures the percentage change in Y for a (small) percentage change in P . Obviously, the two measures are not the same.

ERRORS DUE TO MISSING OBSERVATIONS

Researchers often encounter the problem of MISSING DATA or missing observations. This is quite common in survey-type data and occurs when some respondents do not answer all the survey questions. In time-series data, one can encounter the problem of missing data in several ways. For example, data on some time series are available on a monthly basis (e.g., the unemployment rate), but for some other series the data are available only on a quarterly basis (e.g., those on the Gross Domestic Product, or GDP). Sometimes there are gaps in the data because of special circumstances. For example, during World War II, some data could not be collected.

If we ignore the data on the missing observations and do statistical analysis on whatever data are available, we are likely to obtain estimates of the parameters of interest that are subject to BIAS or are inconsistent. Consider, for example, the following simple regression model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i,$$

where Y is expenditure on leisure travel, X is personal income, and u is the stochastic error term. Suppose data are obtained on 1,000 individuals from a survey that asked them information about travel expenditure in the last 6 months. Suppose further that 100 individuals did not incur any travel expenditure. If one decides to do regression analysis on the basis of the data on 900 individuals only (a method referred to as using CENSORED DATA), it can be shown that the OLS estimators of the β coefficients are biased as well as inconsistent. The problem gets complicated if the income data on the 100 individuals is also unavailable (this is the case of truncation; see CENSORING AND TRUNCATION).

FORECASTING ERRORS

In regression models involving TIME SERIES data, one of the objectives is to use the estimated model to forecast the *future* value(s) of the dependent variables, given the values of the explanatory variables.

Suppose one has developed a regression model to explain the sales of cars in the United States using data, say, for the past 40 quarters. Assume that the model is a good one as judged on the usual statistical criteria. Suppose one wants to use this model to forecast car sales for the next 5 quarters. Assume that the data on the explanatory variables are available for

the 5 quarters. The values of the car sales predicted for the next 5 quarters can be compared with the actual sales data, as information becomes available. No matter how good the model, there is no guarantee that the forecast values and the actual values of car sales will be identical; the difference is the *forecast error*.

A variety of forecasting techniques have been developed, all aiming to reduce the forecast error as far as possible. These range from the simple MOVING AVERAGE type of models to very sophisticated models such as ARIMA (autoregressive integrated moving average) and VECTOR AUTOREGRESSION (VAR). In developing these models, the white noise error term plays a crucial role. To evaluate the performance of the various forecasting models, researchers often use the Akaike information criterion (AIC) or the Schwarz information criterion (SIC). The lower the value of these criteria, the better is the model.

—Damodar N. Gujarati

REFERENCES

- Darnell, A. C. (1995). *A dictionary of econometrics*. Cheltenham, UK: Edward Elgar.
- Diebold, F. X. (2001). *Elements of forecasting* (2nd ed.). Cincinnati, OH: South-Western.
- Gujarati, D. (2002). *Basic econometrics* (4th ed.). New York: McGraw-Hill.

ERROR CORRECTION MODELS

The error correction model (ECM) is a TIME SERIES regression model that is based on the behavioral assumption that two or more time series exhibit an equilibrium relationship that determines both short-run and long-run behavior. The ECM was first popularized in economics by James Davidson, David F. Hendry, Frank Srba, and Stephen Yeo in 1978. In 1987, Robert F. Engle and Clive W. J. Granger demonstrated that cointegrated time series are well represented by an error correction model. Since that time, the ECM has become associated with cointegrated time series. It is, however, important to note that the ECM is a general transformation of an autoregressive distributed lag (ADL) model. Specifically, the ADL model can be rewritten as an ECM such that there are no restrictions on the ADL parameters in the ECM representation. The ECM may thus be used to model equilibrium

relationships involving stationary time series as well as cointegrated time series.

The ECM model is given by

$$\Delta y_t = \lambda_0 + \lambda_1 \Delta x_t - \gamma(y_{t-1} - \beta x_{t-1}) + \varepsilon_t,$$

where $\Delta y_t = y_t - y_{t-1}$, $\Delta x_t = x_t - x_{t-1}$, and γ , the error correction rate, gives the rate at which disequilibrium— $(y_{t-1} - \beta x_{t-1})$ —is corrected. The term in parentheses may include additional independent variables as well. (See Banerjee, Dolado, Galbraith, and Hendry [1993] for a general discussion of ECMs.)

Assume we believe that consumers' sentiment is tied to public views of the president's ability to manage the economy: As people feel more positive about the president's management skills, they also feel more positive about the economy. In other words, we believe that there is a long-run equilibrium relationship between economic approval and consumer sentiment. If news coverage becomes increasingly critical of the president's economic policy such that the public reassesses its view of the president's managerial skill (downward), sentiment may be too high for current economic approval. In this case, the two series may be said to be out of equilibrium, and we would expect the public's economic sentiment to drop. We can capture this relationship in an ADL model or as an ECM. However, if we wish to know how long this adjustment would take—how quickly or slowly economic sentiment would react—we can estimate this effect directly only with the ECM. As parameterized above, the rate of error correction is given by γ . A γ of 0.2 indicates that 20% of the disequilibrium is corrected in the following time period, an additional 20% in the next term period, and so on, until equilibrium is restored. A γ of 0.8 implies a much quicker readjustment. The ECM thus directly estimates the error correction rate. In addition, because the ECM includes both long-run (levels) and short-run (changes) variables on the right-hand side of the model, we can capture the responsiveness of consumer sentiment to short-run changes in evaluations as well as long-run levels of evaluations. The ability to account for how high or low a time series is (how positive or negative evaluations are), as well as the direction it is moving, is an especially attractive feature of the ECM.

In practice, many models of equilibrium relationships are premised on permanent memory and cannot be estimated using a distributed lag model. In this case, the ECM typically is the model of choice. In

addition to the single-equation formulation above, often analysts will rely on the Engle-Granger two-step method for estimating the ECM in the context of cointegration (Engle & Granger, 1987). In this case, the analyst estimates the equilibrium in a first-step regression and enters the first-stage model residuals (the disequilibrium) in the second-stage error correction model.

—Suzanna De Boef

REFERENCES

- Banerjee, A., Dolado, J. J., Galbraith, J., & Hendry, D. F. (1993). *Cointegration, error correction and the econometric analysis of nonstationary series*. New York: Oxford University Press.
- Davidson, J., Hendry, D. F., Srba, F., & Yeo, S. (1978). Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal*, 88, 661–692.
- Engle, R., & Granger, C. W. J. (1987). Co-integration and error-correction: Representation, estimation, and testing. *Econometrica*, 55, 251–276.

ESSENTIALISM

The term *essentialism* is used in philosophy and social science to refer to someone else's allegedly mistaken belief that he or she has accurately described the essential nature of something or someone, or at least that this should be the aim of philosophy or science. The traditional philosophical distinction between the essential natures of things and their appearances feeds into the beginnings of modern social science, with Karl Marx, for example, stressing that science would be unnecessary if things really were as they appeared to be, and that in the study of society the "power of abstraction" replaces procedures such as chemical analysis. Although some commentators commend Marx for essentialism, drawing parallels with Aristotelian philosophy, most see it as an error, whether or not they think he committed it. Those who take the latter view tend to prefer to call his position REALISM.

In recent decades, discussion of essentialism has mostly followed Karl Popper's critique, first advanced in *The Open Society and Its Enemies*, of what he called methodological essentialism: "the view ... that it is the task of pure knowledge or 'science' to discover

and to describe the true nature of things, i.e. their hidden reality or essence" (Popper, 1945/1966, p. 31). To this view, Popper opposes his "fallibilist" position that science is better understood in terms of the refutation of error, and in terms of "verisimilitude" rather than truth.

Popper also upheld a version of realism, but modern realists have mostly argued that his realism is too weak and his critique of essentialism mistaken. Rom Harré (1986, pp. 103–105) and Roy Bhaskar (1978), for example, have defended essences understood in a modest sense, along with notions of natural kinds and natural necessity, in a model of science seen as intrinsically open to refutation and further advance. Other realists, such as Andrew Sayer (1997), have defended realism against the accusation of essentialism and also argued in defense of moderate varieties of essentialism. Technical philosophical discussions of "dispositional essentialism" (Ellis & Lierse, 1994), the claim that a property such as being positively charged is *essentially* linked to a disposition to attract something with a negative charge, are echoed in social science, in Sayer's defense of realist abstraction against "associational thinking" based on regularities that may be merely accidental (Sayer 2000, 2001; see also Holmwood, 2001).

Recent feminist and antiracist critiques of essentialism ("all women/black people are [intrinsically] xyz") have sometimes drawn on deconstructionist or "poststructuralist" positions developed in the critique of essentialism in metaphysics and phenomenology. Other feminists, such as Fuss (1989), however, have criticized the dichotomy between essentialism and social constructionism. These substantive debates therefore intersect with methodological ones; Delanty (1997), for instance, has argued for a synthesis of realist and constructivist or relativist positions in the philosophy of social science.

—R. William Outhwaite

REFERENCES

- Bhaskar, R. (1978). *A realist theory of science*. Hassocks, UK: Harvester Press.
- Delanty, G. (1997). *Social science: Beyond constructivism and realism*. Buckingham, UK: Open University Press.
- Ellis, B. D., & Lierse, C. (1994). Dispositional essentialism. *Australasian Journal of Philosophy*, 72(1), 27–45.
- Fuss, D. (1989). *Essentially speaking*. New York: Routledge.
- Harré, R. (1986). *Varieties of realism*. Oxford, UK: Blackwell.
- Holmwood, J. (2001). Gender and critical realism: A critique of Sayer. *Sociology*, 35, 947–965.

- Popper, K. (1966). *The open society and its enemies* (5th ed., Vol. 1). London: Routledge. (Original work published 1945.)
- Sayer, A. (1997). Essentialism, social constructionism and beyond. *Sociological Review*, 45, 453–487.
- Sayer, A. (2000). System, lifeworld and gender: Associational versus counterfactual thinking. *Sociology*, 34, 707–725.
- Sayer, A. (2001). Reply to Holmwood. *Sociology*, 35, 967–984.

ESTIMATION

One major purpose of statistics is to make an inference about a POPULATION using information on a SAMPLE. Suppose that we are interested in studying the annual income of the adult population in a small American city of 100,000 inhabitants. Then the population would be formed by all the economically active adults in the city at the time of study. If we draw 100 adults randomly from the city, these 100 adults will be our sample. The annual incomes of the adult population in the city can be described by a variety of numerical measures called PARAMETERS. Examples of the parameters include the MEAN annual income of the adult population, the VARIANCE (a measure of variability) of the income, and the ASSOCIATION between the person's education and his or her income. The typical objective of statistical analysis is to make an inference about one or more population parameters. In this article, we consider estimation of population parameters. After a brief discussion of point and interval estimates, we will turn to the ordinary least squares ESTIMATOR and the maximum likelihood estimator. Throughout the discussion, we will use a simple numerical example to illustrate the basic ideas behind these estimators. Almost all textbooks on the subject provide a more general treatment.

POINT AND INTERVAL ESTIMATES

Statistical investigation of the annual income of the adult population in the city could give two types of estimate. The sample mean of \$45,500, for instance, could be given. The sample mean of $\bar{Y} = [\sum_{i=1}^n (Y_i)]/n$ is a single-number or point estimate. We hope that this estimate is not far from the population mean μ . Alternatively, the investigation could present a two-number estimate, for instance, of (\$42,000, \$49,000), meaning that the average annual income of the city population will be likely to fall between \$42,000 and

\$49,000. This alternative form of estimate is called an interval estimate and is designed to enclose the population parameter under study. An interval estimate typically has a higher probability of including the population parameter than the point estimator. Both point and interval estimates have to be produced by an estimator, which is a procedure or method for calculating an estimate from a sample.

THE ORDINARY LEAST SQUARES ESTIMATOR

The ORDINARY LEAST SQUARES (OLS) estimator is often used to estimate the parameters in a LINEAR REGRESSION that is concerned with how one continuous dependent variable Y_i is related to independent variables $x_{1i}, x_{2i}, \dots, x_{ki}$. Such an estimator could take the form

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + e_i. \quad (1)$$

When a linear regression has only one independent variable, it is a simple linear regression of the form

$$Y_i = \beta_0 + \beta_1 x_i + e_i. \quad (2)$$

Such a regression could be used to study how annual income is related to years of education in the city population. Figure 1 is a hypothetical graph that plots annual income against years of education for a sample from the city population. The plot shows a positive relationship between annual income and education. On average, those who have had more years of education tend to have higher incomes. Although many different mathematical functions can be used to model the association between annual income and education, statistical practice usually uses the linear model, such as equation (2). In this particular example, β_0 is the amount of annual income for those who have not received any formal education and β_1 is the amount of annual income associated with one additional year of education. Equation (2) then describes the relationship between annual income and education in the population. A statistical investigation makes use of a sample of the population to calculate $\hat{\beta}_0$ and $\hat{\beta}_1$, which are estimates of the population parameters β_0 and β_1 .

Estimating $\hat{\beta}_0$ and $\hat{\beta}_1$ from the sample amounts to fitting a straight line through the set of data points in Figure 1. We know that no possible line can predict all data points perfectly because education does not predict income perfectly. We want the line to be positioned in such a way that it best represents the set of

data points as a whole. We could try to fit the line by eye, but this works only approximately for a simple linear regression or a two-variable problem. More systematically, we resort to the OLS estimator that yields the fitted line

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \quad (3)$$

The vertical distance between the observed data point y_i and the fitted line represented by $\hat{y}_i$ (or $y_i - \hat{y}_i$) is referred to as deviation or prediction error. The basic idea of the OLS estimator is to select $\hat{\beta}_0$ and $\hat{\beta}_1$ so that we minimize the sum of all the deviations, $\sum_{i=1}^n (y_i - \hat{y}_i)$. However, some data points lie above the fitted line and some lie below the line. The positive errors above the line tend to cancel the negative errors below the line, leaving the sum of all the deviations much smaller than it should be. The standard solution to this problem is to square each error and then minimize the sum of these squared errors (SSE):

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2.$$

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that correspond to the minimum of SSE can be obtained by solving the equations $\partial \text{SSE} / \partial \hat{\beta}_0 = 0$ and $\partial \text{SSE} / \partial \hat{\beta}_1 = 0$. Taking the partial derivatives of SSE with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, and setting the results to zero, we obtain

$$\partial \text{SSE} / \partial \hat{\beta}_0 = -2 \left(\sum_{i=1}^n y_i - n \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i \right) = 0$$

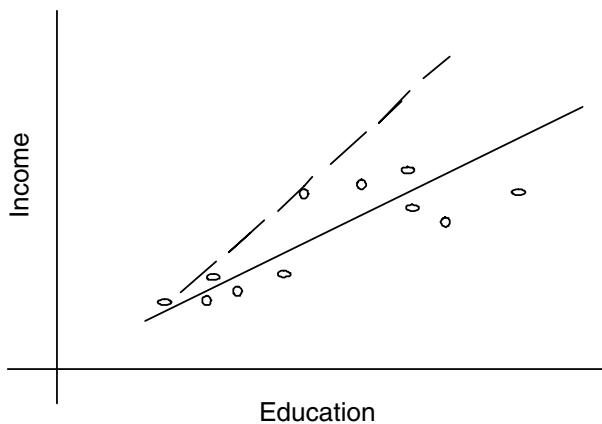


Figure 1 Scatterplot of Income Against Education and the Fitted Line

Table 3 Calculations for the Ordinary Least Squares Fitted Line

Individual i	y_i	x_i	$x_i y_i$	x_i^2
1	35	12	420	144
2	57	18	1026	324
3	26	11	286	121
4	45	12	540	144
5	58	14	812	196
6	40	12	480	144
7	55	16	880	256
8	48	15	720	225
Total	$\sum y_i = 364$	$\sum x_i = 110$	$\sum x_i y_i = 5164$	$\sum x_i^2 = 1554$

and

$$\partial \text{SSE} / \partial \hat{\beta}_1 = -2 \left(\sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \right) = 0.$$

These two equations are the OLS equations and can be solved simultaneously for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Suppose that our sample consists of eight individuals with annual income 35, 57, 26, 45, 58, 40, 55, and 48 (in thousands of dollars) and years of education 12, 18, 11, 12, 14, 12, 16, and 15, respectively. Table 1 illustrates the OLS calculations for the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{y}_i = -7.18 + 3.83x_i. \quad (3)$$

The OLS estimate of the slope thus shows that each year of education is associated with \$3,830 additional annual income.

THE MAXIMUM LIKELIHOOD ESTIMATOR

MAXIMUM LIKELIHOOD ESTIMATOR (MLE) is behind most of the statistical models used in social sciences, such as logistic regression, probit regression, multinomial regression, Poisson regression, negative binomial regression, a variety of event history models, growth curve models, and multilevel linear models. The EM algorithm, which gained popularity over the past 15 years, is essentially a computing method for

maximum likelihood estimation. Even the parameter estimates of the OLS regression discussed earlier can be considered the estimates of the MLE under the assumption that the data have a normal distribution.

Suppose that we want to know the mean annual income of the adult population in our small city. A feasible way of estimating this is to draw and work with a random sample. Suppose that our sample consists of eight individuals with incomes as specified earlier (the MLE should really apply to large samples; the small sample of eight is meant to simplify the illustration). To proceed with the MLE, we must specify the DISTRIBUTION of the data. By far the most assumed distribution for linear data is the NORMAL DISTRIBUTION. Thus, we assume that each income of the eight individuals comes from a normal distribution: $Y_1 \sim N(\mu, \sigma^2), \dots, Y_8 \sim N(\mu, \sigma^2)$ with density

$$f(y_i) = 1/\sqrt{2\pi\sigma^2} e^{-(y_i - \mu)^2/2\sigma^2}.$$

The parameter μ is not simply the average annual income. Rather, it is the mean of the normal distribution representing the annual income in the population.

The next step in the MLE is to write down the likelihood function of the data. The likelihood function for individual i is the density for the individual viewed as a function of the unknown parameters:

$$L_i(\mu; y_i) = 1/\sqrt{2\pi\sigma^2} e^{-(y_i - \mu)^2/2\sigma^2}.$$

Assuming that the individual observations in the sample are independent, we can write down the joint likelihood function for the whole sample as

$$L(\mu; y_1, y_2, \dots, y_8) = \prod_{i=1}^8 L_i = \left(1/\sqrt{2\pi\sigma^2}\right)^8 e^{-[\sum_{i=1}^8 (y_i - \mu)^2/2\sigma^2]}.$$

The MLE is that value of the μ that maximizes the joint likelihood function $L(\mu, y_1, y_2, \dots, y_8)$. In practice, however, it is much easier to maximize $\ln L(\mu)$, the natural logarithm of $L(\mu)$, than to maximize $L(\mu)$. The quantities associated with the likelihood function $L(\mu)$ often become too small or large for computing. The justification for this is that whatever maximizes the log likelihood must also maximize the likelihood. With the logarithm, $L(\mu)$ is simplified to

$$\ln L(\mu) = 8 \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{\sum_{i=1}^8 (y_i - \mu)^2}{2\sigma^2}.$$

Finding the MLE value requires taking the partial derivative of the log likelihood function with respect to μ , setting the derivative equal to zero, and then solving the equation for the unknown parameter:

$$\begin{aligned} \frac{\partial \ln L(\mu)}{\partial \mu} &= \frac{\sum_{i=1}^8 2(y_i - \mu)}{2\sigma^2} = 0 \\ \hat{\mu} &= \sum_{i=1}^8 y_i / 8 = \bar{y} \\ &= \frac{35 + 57 + 26 + 45 + 58 + 40 + 55 + 48}{8} \\ &= 45.5. \end{aligned}$$

Using this result, we can obtain the MLE for σ^2 in a similar way:

$$\begin{aligned} \frac{\partial \ln L(\mu)}{\partial \sigma^2} &= \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^8 (y_i - \mu)^2 = 0, \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^8 (y_i - \bar{y})^2. \end{aligned}$$

Thus, the maximum likelihood estimators of the mean and the variance of the normal distribution turn out to be the corresponding sample characteristics. In simple models, the MLE is often the “obvious” estimator; however, the obviousness usually is lost when the model becomes slightly more complicated. The obviousness of the MLE in the simplest cases may be considered reassuring because we probably want to see the MLE in the simplest cases as simple and easily understood.

Maximum likelihood estimation is a method with an intuitive appeal. Apart from the obviousness of MLE in the simplest cases, the ML estimator is the population estimate that is most likely to produce the observed sample. More fundamentally, the popularity of MLE is due to many of the attractive ASYMPTOTIC PROPERTIES (large sample properties) that it possesses. Under broad conditions, an ML estimate is consistent, efficient, and normally distributed in its sampling distribution, with clearly defined and easily computed mean and variance.

The estimator $\hat{\theta}_n$ based on a sample of size n is said to be a consistent estimator of the population parameter θ if for any positive number ε , $\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$. The maximum likelihood estimator is consistent. This result says that the estimates of an ML estimator would get closer and closer to the

population parameter as the sample size gets larger and larger and that the estimates would converge to the population parameter when the sample size approaches positive infinity. The consistency result is intuitive. Let's consider coin tossing. The coin is tossed n times with the unknown probability (p) of resulting in a head. If the tosses are independent, the number of heads Y in n tosses has a binomial distribution. The sample proportion Y/n is an estimator of p . It is not difficult to understand that as the number of tosses n increases, our estimates of p would get closer to the actual value of p , and that as the number of tosses approaches positive infinity, our estimates would converge to the actual p .

The efficiency of an estimator is indicated by the amount of variability associated with the estimates of the estimator. A population parameter θ may be estimated by two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$. The two estimators represent two different mathematical procedures. Given a sample, the two procedures would generate two estimates. Given a large number of samples of size n , each procedure would generate a set of estimates. Suppose both sets of estimates have a mean equal to θ , meaning that both $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased. Then if the variance of the estimates associated with $\hat{\theta}_1$ is smaller than the variance of the estimates associated with $\hat{\theta}_2$, we say that the estimator $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$. The well-known result by Rao-Cramér provides a lower bound for every unbiased estimator of θ . If an unbiased estimator has a variance equal to the Rao-Cramér lower bound, we know that no unbiased estimator has a smaller variance and that the estimator under consideration is the unbiased minimum variance estimator. Under fairly general conditions, it can be proved that the ML estimator $\hat{\theta}$ of θ has an approximate normal distribution with mean θ and a variance equal to the Rao-Cramér lower bound. Therefore, at least approximately $\hat{\theta}$ is an unbiased minimum variance estimator. Therefore, the ML estimator can be considered an efficient estimator in the sense that it will make good use of the sample collected.

The last property of the ML estimator we consider here is perhaps the most useful. In most regular cases, it can be shown that the ML estimator based on large samples has a normal distribution with mean θ and the variance $[I(\theta)]^{-1}$, where

$$I(\theta)^{-1} = \left(-E \left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right] \right)^{-1}$$

when θ has a single parameter.

The variance can be evaluated at $\hat{\theta}$ when the expected value is available. In practice, the expected value of the second derivative of the log likelihood will seldom be available. In such cases, the empirical information matrix rather than the information is often computed:

$$\left(-\frac{\partial^2 \ln L(\hat{\theta})}{\partial \hat{\theta}^2} \right)^{-1}.$$

This result holds whether or not the data have a normal distribution. This result is the basis of the z -test in regression analysis that is so prevalent in social science research.

BAYESIAN ESTIMATION

Bayesian estimation has found much use in recent years in the mainstream statistical and biostatistical literature. A classical estimator such as the ordinary least squares estimator or the maximum likelihood estimator is determined entirely by the sample. The Bayesian estimation, in contrast, draws information from both the sample and the PRIOR DISTRIBUTION. When additional data become available, one updates one's prior distribution. The Bayesian estimation incorporates one's prior knowledge on the problem and is more subjective. For large samples and regular models, the prior distribution usually is overwhelmed by the amount of data and the Bayesian estimation yields the same estimates as does maximum likelihood estimation. Thus, Bayesian estimation is not a general substitute for maximum likelihood estimation.

There are, however, at least two situations in which Bayesian estimation can play an important role. The first is when the number of observations is small and the information about each observation is abundant. Studies in macrosociology sometimes belong to this category. When a sample consists of countries, the number of countries in the sample is necessarily limited, but the background information about each country can be huge. In such cases, taking advantage of prior knowledge about each country—that is, information outside the observed data—may contribute to the analysis. The second situation is when the models are too difficult to estimate by classical statistical approaches. Examples include nonlinear multilevel models. The log likelihood of these models frequently is involved, with multiple integrals that do not have a closed expression. MARKOV CHAIN MONTE CARLO (MCMC) METHODS

developed within the Bayesian perspective bypass the problem by drawing samples from the required distribution and then describing the sample averages. There are other estimation methods not discussed here, for example, the method of moment estimation and indirect estimation techniques for demographic data.

—Guang Guo

REFERENCES

- Greene, W. H. (1997). *Econometric analysis* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.
- Hogg, R. V., & Craig, A. T. (1978). *Introduction to mathematical statistics* (4th ed.). New York: Macmillan.
- Lee, P. M. (1997). *Bayesian statistics: An introduction* (2nd ed.). London: Arnold.

ESTIMATOR

An estimator, also known as a (sample) STATISTIC, is a rule, method, formula, or procedure that tells how one can estimate a POPULATION quantity, such as the population mean, population variance, or population CORRELATION coefficient. It generally is expressed as a function of sample values. The numerical value taken by an estimator in a given sample is known as an estimate. Because its value will differ from sample to sample, an estimator is a RANDOM VARIABLE and will have a PROBABILITY or SAMPLING DISTRIBUTION. The sampling distribution of an estimator is of great importance in assessing the reliability of the estimate in relation to its population value. All of this can be explained as follows.

Let X be a random variable with a PROBABILITY DENSITY FUNCTION (PDF) of $f(X)$. For simplicity of exposition, assume that this PDF has two PARAMETERS, μ_x (population mean) and σ_x^2 (population variance). Based on a RANDOM SAMPLING of n observations from this PDF, suppose one wants to estimate the mean value, μ_x . The most commonly used estimator of the population mean is the SAMPLE MEAN, defined as $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Thus $\bar{X}$ is a rule or formula that says that to obtain the sample mean value, one simply adds all the values of X in the sample and divides the sum by n , the number of observations in the sample. Thus, $\bar{X}$ is an estimator of μ_x . (Besides the sample mean, the sample MODE and MEDIAN are the other common estimators of CENTRAL TENDENCY.)

If one is interested in estimating the population variance, the most commonly used estimator is the SAMPLE VARIANCE, denoted by S_x^2 , which is defined as follows:

$$S_x^2 = \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2.$$

According to this formula, to estimate the sample variance, subtract the mean value of X from each individual value of X , square the difference, sum these squared differences, and divide the sum by DEGREES OF FREEDOM ($n-1$). The sample variance thus defined is an UNBIASED estimator of the true σ_x^2 .

The formulas for computing the sample mean and sample variance just given provide what are called the POINT ESTIMATES. That is, for a given sample they will provide a single (point) estimate. Because an estimator is a random variable, as its value will vary from sample to sample, how reliable is a point estimate of its true population value? In statistics, the reliability of a point estimator is measured by its STANDARD ERROR. Therefore, instead of relying on the point estimate alone, one may construct an interval around the point estimator, say within two or three standard errors on either side of the point estimator, such that this interval has, say, a 95% probability of including the true parameter value. The interval thus established provides the so-called *interval estimator*.

Knowing that the variance of X is σ_x^2 , statistical theory shows that the variance of

$$\bar{X} = \frac{\sigma_x^2}{n}.$$

Therefore, the standard error of $\bar{X}$, which is the square root of the variance of $\bar{X}$, is

$$\frac{\sigma_x}{\sqrt{n}}.$$

Statistical theory also shows that if the number of observations is fairly large, the sample mean $\bar{X}$ approaches the NORMAL DISTRIBUTION with mean μ_x and variance $\frac{\sigma_x^2}{n}$ (this is as per the CENTRAL LIMIT THEOREM). If the random variable X is from a normal population to begin with, $\bar{X}$ is normally distributed with the stated parameters regardless of the sample size. In other words, for large samples,

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right);$$

that is, $\bar{X}$ is normally distributed with the stated mean and variance. Thus, the probability or sampling distribution of the sample mean is normal.

Using the properties of the normal distribution, it then follows that the interval

$$\left(\bar{X} \pm 1.96 \frac{\sigma_x}{\sqrt{n}} \right)$$

provides a 95% confidence interval for the population mean. If 100 intervals like this are established, chances are that 95 out of 100 of them will include the true μ_x . The preceding interval is thus an interval estimator of the true mean.

There are several methods of obtaining a point estimator, the prominent ones being the method of ORDINARY LEAST SQUARES (OLS) and the method of MAXIMUM LIKELIHOOD ESTIMATION (ML). Of the two, the method of OLS is mathematically simpler to use. To illustrate this, consider the following MULTIPLE REGRESSION:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i,$$

where Y is the dependent variable, the X s are the explanatory variables, and u_i is the stochastic error term. To estimate the model, given the data, OLS proceeds as follows: Rewriting the preceding equation, we obtain:

$$u_i = Y_i - \beta_1 - \beta_2 X_{2i} - \beta_3 X_{3i} - \cdots - \beta_k X_{ki}.$$

Thus, u_i represents the difference (or error) between the actual Y values and their estimated values from the multiple regression. Instead of minimizing the sum of these errors, OLS minimizes the squares of these errors. That is, it minimizes

$$\sum u_i^2 = \sum (Y_i - \beta_1 - \beta_2 X_{2i} - \beta_3 X_{3i} - \cdots - \beta_k X_{ki})^2.$$

Because the values of Y and the X variables will be provided by the sample at hand, the preceding error sum of squares will depend on the values taken by the β coefficients. OLS tries to find those values of the β coefficients that will make $\sum u_i^2$ as small as possible. The actual algebraic formulas to compute the k beta coefficients are rather tedious. Using matrix algebra, the OLS estimators can be shown as follows:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

where $\hat{\beta}$ is a $(k \times 1)$ vector of the k beta coefficients, $\mathbf{X}$ is an $(n \times k)$ data (or observation) matrix, and $\mathbf{y}$ is

an $(n \times 1)$ vector of the observations on the dependent variable; the $\hat{\cdot}$, called a hat or cap, denotes an estimator. Although β (vector) is unknown but fixed, $\hat{\beta}$ (vector) is random because its value will depend on the sample-specific values of the Y and the X variables. It can be further shown that the OLS estimator of the error variance is given by

$$\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n - k},$$

where $\hat{u}_i$ is the estimated error term and where $(n - k)$ represents the number of degrees of freedom, which is equal to the sample size n minus the number of parameters estimated in the model.

In choosing among competing estimating methods, one can consider the statistical properties of the estimators obtained from the various methods. The commonly used statistical properties are linearity, UNBIASEDNESS, minimum variance, EFFICIENCY, sufficiency, asymptotic unbiasedness (see ASYMPTOTIC PROPERTIES), asymptotic efficiency, and consistency. Not every estimation method will satisfy all these statistical properties, especially if the sample size is small.

For example, if we assume that the multiple regression considered above satisfies the assumptions of the classical LINEAR REGRESSION model, it can be proved that the OLS estimators are linear, are unbiased, and have minimum variance in the class of linear unbiased estimators; they are often described as BEST LINEAR UNBIASED ESTIMATORS (BLUE). (The classical linear regression model assumes that the values taken by the explanatory variables are fixed in repeated sampling, that the error terms u_i are not correlated, and the error variance of u_i , σ_u^2 , is constant or HOMOSKEDASTIC.) It can further be shown that the OLS estimator of error variance given above is an unbiased estimator of the true σ_u^2 .

If, in addition to the assumptions of the classical linear regression model being met, it is assumed that the error term u_i is normally distributed, then it can be shown that the maximum likelihood (ML) estimators of the β coefficients are identical to the OLS estimators. However, the ML estimator of the error variance of u_i is

$$\sigma_u^2 = \frac{\sum \hat{u}_i^2}{n},$$

which is different from the OLS estimator given above. As noted previously, the OLS estimator of the

error variance is unbiased, which means that the ML estimator of the error variance is biased. The difference between the two estimators is due to the fact that the OLS estimator takes into account the degrees of freedom, whereas the ML estimator does not. Of course, as the sample size n increases indefinitely, the two estimators will converge. In other words, asymptotically the ML estimator of the error variance is unbiased.

—Damodar N. Gujarati

REFERENCES

- Gujarati, D. N. (2002). *Basic econometrics* (4th ed.). New York: McGraw-Hill.
- Kendall, M. G., & Buckland, W. R. (1971). *A dictionary of statistical terms* (3rd ed.). Edinburgh: Oliver & Boyd.
- Vogt, P. W. (1993). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences*. Newbury Park, CA: Sage.

ETA

In the social sciences, the term eta (Greek symbol η) probably is best known as it is used by Hays (1988) and Kirk (1995). The term does not have any particular meaning in the context of basic statistical training or theory. For example, eta is not found in the dictionary of statistical terms by Kendall and Buckland (1982). A search of all published journal articles in the *Current Index to Statistics*, for the years 1980 through 2002, based on the keyword eta (or eta-squared), turned up zero papers. Hays and Kirk use the term in the context of the ANALYSIS OF VARIANCE, but several recent texts devoted to this topic include no mention of eta. Some use the term eta to refer to a measure of effect size that is a function of the Student's t -statistic, and it has played a role in structural equation models. Sometimes eta-squared is used in regression and refers to the COEFFICIENT OF multiple DETERMINATION, which is more commonly labeled R -SQUARED.

Consider a standard ANALYSIS OF VARIANCE (ANOVA) model where the means of J independent groups are to be compared assuming normality and that all J groups have a common variance. The term eta-squared refers to a measure of association intended to reflect any nonlinear correlation between an independent and a dependent variable. For example, imagine that five different amounts of some medication

are under investigation regarding their ability to treat depression. For illustrative purposes, assume that 10, 20, 30, 40, and 50 units of the drug are being studied. Of interest is the mean of some outcome measure that reflects depression. If the means of the five groups differ, one possibility is that with increasing amounts of medication, the means increase in a linear fashion. Another possibility is a nonlinear association in which, for example, the means increase by the same amount when increasing the amount of the drug from 10 to 20, and from 20 to 30, but then the means decrease when moving from 30 to 40 and 40 to 50. The term eta-squared refers to a type of correlation intended to reflect this, or any, nonlinear association. Generally, it is meant to reflect any departure from a linear trend.

In ANOVA, two standard quantities are the SUM OF SQUARES between groups (SSBG), which reflects the variation among the sample means, and the total sum of squares (SSTOT). The estimate of eta is the square root of SSBG divided by SSTOT. There is also a quantity that reflects the linear association that is based on a particular linear contrast of the means; it is essentially a variation of Pearson's correlation, r . If there is a linear association, then eta and r should have similar values, so a simple check for a nonlinear association is to compare these two quantities. A detailed illustration can be found in Kirk (1995, pp. 197–198).

—Rand R. Wilcox

REFERENCES

- Hays, W. L. (1988). *Statistics*. Fort Worth, TX: Holt, Rinehart and Winston.
- Kendall, M. G., & Buckland, W. R. (1982). *A dictionary of statistical terms*. New York: Longman.
- Kirk, R. E. (1995). *Experimental design*. Pacific Grove, CA: Brooks/Cole.

ETHICAL CODES

A code of ethics specifies the proper conduct of members of a particular group. This article focuses on codes of ethics in the social sciences. Codes are controversial. They ignore the duty of a professional to simply be moral and are rarely used for guidance by practicing professionals. Codes testify to the claim that the profession recognizes special obligations to society that transcend economic self-interest or normal

standards of morality. They help moral professionals resist pressures to compromise as well as legitimizing one's objections to poor work by other professionals and enhancing the profession's reputation and working environment. Professional societies may use them to censure or expel misbehaving members. Violation of ethics codes is not, per se, legally punishable, but it may be admissible as evidence in some legal proceedings. The role of codes is best understood in historical context.

HISTORY

Until about 1800, professionals were considered gentlemen who needed no written rules of behavior. Mere suggestion that a professional was dishonorable might result in law suits, pamphlet wars, or even duels. The inappropriateness of such idiosyncratic standards was recognized in 1792, when physicians at the Manchester Infirmary waged a pamphlet war and work stoppage during the outbreak of an epidemic. Outraged hospital trustees requested that Dr. Thomas Percival draft rules to prevent future breakdowns of professional morality. Percival's rules, published in 1794, listed specific duties of physicians, asserting the importance of common professional standards and establishing the moral independence of physicians over authorities who might pressure them to lower standards of medical care. In 1847, the American Medical Association was the first national professional society anywhere to adopt Percival's rules and call them a "code of ethics."

GROWTH AND DEVELOPMENT OF CODES IN THE SOCIAL SCIENCES

The American Psychological Association published the first social science code in 1953, following the now-familiar practice of convening a committee, gathering descriptions of ethical dilemmas encountered by members, and developing, through consensus, appropriate rules. Currently, there are codes of ethics for every branch of social science, as well as for subareas within them, such as public opinion research, and clinical hypnosis. Changes within society (e.g., emphasis on privacy), regulations (e.g., institutional review boards), professional roles, and research topics require frequent review and revision of codes. Typically, a professional association's ethics committee publishes articles describing emerging issues,

suggests approaches to resolving them, and seeks membership input. Drafts of the emerging code are revised and published until membership consensus is reached.

USING CODES

Because codes of ethics are created in response to anticipated ethical conflicts, they are best understood and interpreted in the context of real-life ethical ambiguity. Case studies of ethical dilemmas are effective instructional tools for teaching the use of codes and are now an important part of each discipline's professional literature.

A voluminous literature on codes of ethics now exists. Collections of codes, annotated bibliographies, casebooks, research guidelines, methodology integrating tenets of valid research and ethics, as well as guidance in authoring codes and teaching their use, can be found at the URL of the Center for the Study of Ethics in the Professions, Illinois Institute of Technology (www.iit.edu/departments/csep/PublicWWW/codes/). The archives on codes of ethics, including their history, construction, use, and examples across the professions, are housed at the same Center for the Study of Ethics in the Professions.

—Joan E. Sieber

ETHICAL PRINCIPLES

Ethical discourse is conducted at various levels of abstraction. Moving from higher to lower levels of abstraction, there are ethical theories, principles, rules or norms, codes, and judgments. The nature and role of ethical principles is best understood in relation to other elements in this hierarchy.

Judgments are conclusions about a particular action. An example is the following:

I don't think we should tell the 5-year-olds in our study of resistance to temptation that we observe whether they steal the money we leave around, because children that age readily take attractive items without realizing that they are doing something wrong. Rather, we should chat with them and make sure they feel OK about their actions and make sure their parents understand relevant aspects of child development and of our study.

This very concrete and specific judgment can be justified on the basis of more abstract and general rules, principles, and theories.

A *code of ethics* specifies the proper conduct of members of a particular group (see ETHICAL CODES). That is, a code is a set of rules or norms of a particular group. The code recognizes the group's special obligations to society—obligations that transcend normal standards of morality. For example, the code of ethics of the American Psychological Association requires appropriate debriefing and also justifies the judgment and norm discussed above concerning debriefing of young participants in deception research. It states that

(a) Psychologists provide a prompt opportunity for participants to obtain appropriate information about the nature, results, and conclusions of the research, and psychologists attempt to correct any misconceptions participants may have.

(b) If scientific or humane values justify delaying or withholding this information, psychologists take reasonable measures to reduce the risk of harm. (American Psychological Association, 1992, p. 1204)

Note that this statement is more generalized than the rule or norm stated above. It invites judgment about appropriate interpretation of the code.

Rules or norms often are informal and unwritten but generally are understood as general statements concerning what is right and wrong. For example, a researcher might justify the above judgment with the following rule or norm: "Researchers should debrief subjects, but should not do so in a way that would unnecessarily harm or worry them or their parents."

A *principle* is an even more generalized rule, such as "respect research participants," "be fair," "prevent harm," and "promote benefit." It states ideals, ignores conflicts that may arise between principles, and gives no instructions as to *how* one is to carry out these ideals. For example, the American Psychological Association sets forth Principle F as follows:

Psychologists are aware of their professional and scientific responsibilities to the community and the society in which they work and live. They apply and make public their knowledge of psychology in order to contribute to human welfare. Psychologists are concerned about and work to mitigate the causes of human suffering. When

undertaking research, they strive to advance human welfare and the science of psychology. Psychologists try to avoid misuse of their work. (American Psychological Association, 1992, p. 1600)

Ethical theories are systematically related bodies of principles and rules. The following are highly simplified descriptions of some major types of ethical theories. *Utilitarian* (consequential) theories hold that one should try to do whatever promises to maximize happiness in the world. *Deontological* ethics (Immanuel Kant's version) holds that one has specific duties that must be carried out, quite apart from their supposed consequences, because these duties are always morally right unless superseded by some higher principle such as saving a life. These might include duties such as truth-telling or treating people as ends in themselves rather than as means to ends. Similarly, some deontologists might hold that certain acts, such as killing a fetus, are always wrong. *Theist* ethics hold that one should do what will please God.

As these simple definitions and examples illustrate, different ethical theories would justify different principles, rules, and judgments; hence, ethicists may disagree about what is a morally or ethically correct course of action. Additionally, the same theories and principles may be interpreted differently by different persons or in different contexts.

Disagreements, exceptions, and conflicts. There are always exceptions to ethical theories, principles, and rules. For example, regarding exceptions to theories, deontologists who hold that killing another person is always wrong might argue that there is such a thing as a "just war" in which killing is justified, and deontologists who require truth-telling might permit lying to save an innocent person. Ethical principles are often found to be in conflict with one another or subject to exceptions. A major role of ethical theories is to help resolve conflicts between principles. For example, the hypothetical investigator who studied resistance to temptation by 5-year-olds reconciled the requirement to debrief honestly with the requirement not to cause harm, looking to the utilitarian principle of beneficence—the researcher's obligation to maximize possible benefit and minimize possible harm—for justification.

As this introduction to ethical principles and their place in the hierarchy of ethical discourse is intended to convey, there are both many ways to talk about

ethics and many different positions one might take. Seemingly incomprehensible differences of opinion often can be understood by identifying the kind of theory on which opinions are based. For example, persons who are horrified at research practices that involve some degree of concealment or deception seem to be taking a deontological position. Thus, cultural anthropologists who observe the behavior of another without revealing the critical stance that will be taken in the scientific discussion of that behavior might be regarded as doing wrong because they are treating subjects as means to the end of learning about human behavior, and not as ends in themselves. A utilitarian, on the other hand, might argue that humankind is much better off understanding itself, even at the cost of minor deception of individual subjects; the end justifies the means. The ethical principles set forth under the *Belmont Report* and promulgated via the Federal Regulations for Protection of Human Subjects of Research are a mixture of utilitarian and deontological ethics.

One of the most difficult ethical problems for investigators to resolve—as they seek to conduct risky research that promises to yield highly important knowledge—arises when that research also requires some compromise of a deontological principle by resorting to a utilitarian principle. For example, research on minors who inject drugs may promise great benefit for society because it may further society's ability to curtail such drug use. However, it may be impossible for the researcher to obtain from parents their permission to conduct the research; the parents may be unaware of their child's drug use or may be dependent on their child's drug dealing for income.

Ethical principles have the advantage of being broad and general, hence of providing flexible guidelines for considering a wide range of issues. By the same token, they are vague on specifics and may be easily misinterpreted. With this proviso in mind, we turn to the "Belmont Principles" set forth by the U.S. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (National Commission) in 1978. The National Commission had, as part of its charge, the mission to set forth basic ethical principles that should govern human research. After much deliberation, it set forth three principles that support six generally agreed upon norms of scientific research. The ethical principles are as follows:

a. *Beneficence*—maximizing good outcomes for science, humanity, and the individual research

participants while avoiding or minimizing unnecessary risk, harm, or wrong

b. *Respect*—protecting the autonomy of (autonomous) persons, with courtesy and respect for individuals as persons, including those who are not autonomous (e.g., infants, mentally retarded persons, or persons afflicted with senility)

c. *Justice*—ensuring reasonable, non-exploitative, and fully considered procedures and their fair administration, along with fair distribution of costs and benefits among persons and groups (i.e., those who bear the risks of research should be those who benefit from it)

The six norms of scientific behavior, with letters in parentheses indicating the principles on which each norm is based, are as follow:

1. *Valid research design*. Valid research takes account of relevant prior findings, methods, and theory. Invalid research yields useless results, disrespects subjects, and wastes resources (A, B).

2. *Competence of researcher*. The researcher must be capable of carrying out research validly (A, B).

3. *Identification of consequences*. Risks and benefits should be identified. Procedures must be adjusted to respect privacy, ensure confidentiality, maximize benefit, minimize risk, and meet other goals (A, B, C).

4. *Selection of subjects*. Subjects must be appropriate to the study, representative of the population that will benefit, and appropriate in number (A, B, C).

5. *Voluntary informed consent*. Voluntary informed consent should be obtained (A, B, C).

6. *Compensation for injury*. The investigator is responsible for harm to subjects. Subjects must be informed whether harm will be compensated (A, B, C).

The Belmont Principles reflect the individualism and pragmatism of American culture. A more paternalistic or communitarian culture might place less emphasis on respect for autonomy and greater emphasis on beneficence. A more theocratic culture might discourage research on humans if it held that all behavior is according to God's will. Aside from such extreme moral viewpoints, the National Commission might have chosen a somewhat different set of fundamental principles. Many have observed that the current

emphasis on respect for autonomy leaves many research participants with decisions they are ill-equipped or unwilling to make. Consequently, many subjects consent or refuse to participate without understanding (or even reading) the informed consent statement. Would a set of principles based more on caring or professional responsibility of the investigator to do no harm be more ethical? There are no easy answers.

—Joan E. Sieber

REFERENCES

- American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist*, 47, 1597–1611.
- National Commission for Protection of Human Subjects of Biomedical and Behavioral Research. (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research* (DHEW Publication No. (OS) 78–0012). Washington, DC: Government Printing Office.
- Solomon, W. D. (1978). Ethics: Rules and principles. In W. T. Reich (Ed.), *Encyclopedia of bioethics*. New York: Free Press.

ETHNOGRAPH

Ethnograph is a software program designed for computer-assisted qualitative data analysis. As such, it allows qualitative data to be coded and retrieved. It also supports GROUNDED THEORY practices, such as the generation of MEMOS.

—Alan Bryman

ETHNOGRAPHIC CONTENT ANALYSIS

Ethnographic content analysis (ECA) refers to an integrated method, procedure, and technique for locating, identifying, retrieving, and analyzing documents for their relevance, significance, and meaning (Altheide, 1987, 1996). The emphasis is on discovery and description, including search for contexts, underlying meanings, patterns, and processes, rather than mere quantity or numerical relationships between two or more variables (Altheide, 1996).

A document is defined as any symbolic representation and meaning that can be recorded and/or retrieved for analysis. Document analysis will expand as recording technologies improve and become more accessible. These technologies include those of print and electronic media, audio tapes, visuals (e.g., photos, home videos), clothing/fashion, Internet materials, information bases (e.g., Lexis/Nexis), and fieldnotes.

ECA or qualitative document analysis involves emergent and theoretical sampling (Glaser & Strauss 1967) of documents from information bases (including those developed by a researcher, such as FIELDNOTES), development of a protocol for more systematic analysis, and then comparisons to clarify themes, frames, and discourse. For example, if one is interested in studying TV violence, it is not an act of violence per se that is socially significant, but rather how that act is linked to a *course of action* or a scenario as part of an *entertainment emphasis* (e.g., “bad guys get shot by good guys in order to achieve justice”). Alternatively, the scenario might be that the use of violence is somehow linked to bravery, cunning, skill, or (of course) sex. The latter are themes or general messages that are reiterated in specific scenarios. The aim, then, is to query how behavior and events are placed in context, and what themes, frames, and discourse are being presented. The basic steps include:

- Pursuing a specific problem to be investigated
- Becoming familiar with the process and context of the information source (e.g., ethnographic studies of newspapers or television stations) and exploring possible sources (perhaps documents) of information
- Becoming familiar with several (6–10) examples of relevant documents, noting particularly the format, and selecting a unit of analysis (e.g., each article), recognizing that the unit of analysis may change
- Listing several items or categories (variables) to guide data collection and drafting a protocol (data collection sheet)
- Testing the protocol by collecting data from several documents
- Revising the protocol and selecting several additional cases to further refine the protocol

A dynamic use of ECA involves “tracking discourse,” or following certain issues, words, themes,

and frames over a period of time, across different issues, and across different news media. Initial manifest coding incorporates emergent coding and theoretical sampling in order to monitor changes in coverage and emphasis over time and across topics. For example, in a study of “fear” (Altheide, 2002), a protocol was constructed to obtain data about date, location, author, format, topic, sources, theme, emphasis, and grammatical use of fear (as noun, verb, or adverb). The contexts for using the word “fear” were clarified through theoretical sampling and CONSTANT COMPARISON to delineate patterns and thematic emphases. Materials were enumerated, charted, and analyzed qualitatively using a word processor and a qualitative data analysis program—NUD*IST—as well as quantitatively.

—David L. Altheide

REFERENCES

- Altheide, D. L. (1987). Ethnographic content analysis. *Qualitative Sociology*, 10, 65–77.
- Altheide, D. L. (1996). *Qualitative media analysis*. Newbury Park, CA: Sage.
- Altheide, D. L. (2002). *Creating fear: News and the construction of crisis*. Hawthorne, NY: Aldine de Gruyter.
- Glaser, B. G., & Strauss, A. L. (1967). *Discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

ETHNOGRAPHIC REALISM

Ethnographic realism holds that ETHNOGRAPHY can and should strive to document sociocultural structures, processes, and situations as existing independently of the researcher.

For much of its history, ethnography has been committed to REALISM and to methods of investigation based on that commitment. For example, anthropologists have long argued that other societies must be studied through close PARTICIPANT OBSERVATION in order for their culture and social organization to be understood properly. Within 20th-century U.S. sociology, Herbert Blumer and others recommended much the same approach, emphasizing the need to capture the interpretative and processual character of human social life, as well as criticizing methods that failed to do this. In the late 1960s, this realism—sometimes also referred to as NATURALISM—was summarized by

David Matza as the obligation to remain true to the nature of the phenomenon under study.

In the 1980s and 1990s, however, ethnographic realism came under challenge. The critics raised questions about the very possibility of providing accounts of the “nature” of social phenomena, sometimes denying that these phenomena exist independently of the process of inquiry. This critique arose, to some extent, out of ethnography itself. Typically, ethnographers have portrayed people as actively making sense of the world, thereby developing distinctive cultures. If this approach is applied reflexively to ethnographic accounts themselves, the conclusion may be reached that these accounts are creative constructions that reflect the sociocultural identities of researchers, rather than being objective representations. The tendency to draw this conclusion was encouraged by the influence of some strands of continental philosophy, notably PHENOMENOLOGY, HERMENEUTICS, and POSTSTRUCTURALISM (though these positions do not all imply epistemological RELATIVISM or skepticism).

Another important component of the critique pointed out that ethnographic accounts often rely on the same literary devices that are used in realist fiction. The critics insisted that no form of writing is merely representational: All language use is constitutive and performative, and no writer should pretend otherwise.

Along with criticisms of ethnographic realism on epistemological and rhetorical grounds, there has also been an ethical and political critique. Drawing on radical Leftist challenges to sociology in the 1960s, as well as on feminism, this critique has argued that realist ethnography amounts to voyeurism and/or serves as a means of surveillance by which those in subordinate social positions are kept under control.

In line with these arguments, some critics of ethnographic realism have sought to develop new forms of writing designed to subvert any impression that ethnographic texts can provide objective accounts of the world, to display their own constructed and indeterminate character, and/or to give voice to those who are seen as being “silenced” by mainstream accounts (see Denzin, 1997).

There has been some defense of ethnographic realism against these criticisms. The self-undermining character of skeptical and relativist arguments has been reiterated. Furthermore, a distinction has been drawn between the naive realism that is sometimes ascribed to ethnographers of the past and more

sophisticated forms. The latter take account of epistemological conundrums and of the socially contextualized character of all research but do not deny the possibility of social scientific knowledge or insist that ethnography is inevitably implicated in political struggle supporting or resisting the status quo (see Hammersley, 1992).

—Martyn Hammersley

REFERENCES

- Atkinson, P. (1990). *The ethnographic imagination: Textual constructions of reality*. London: Routledge.
- Denzin, N. K. (1997). *Interpretive ethnography*. Thousand Oaks, CA: Sage.
- Hammersley, M. (1992). *What's wrong with ethnography?* London: Routledge.
- Matza, D. (1969). *Becoming deviant*. Englewood Cliffs, NJ: Prentice-Hall.

ETHNOGRAPHIC TALES

ETHNOGRAPHY, a long-established data collection method in anthropology, provides rich, detailed information about individuals and groups. As a qualitative research method, ethnography typically involves PARTICIPANT or NON-PARTICIPANT OBSERVATION and/or IN-DEPTH INTERVIEWS, allowing researchers to get a comprehensive look at the population of interest.

A good ethnographic account provides “thick description,” or a detailed account of the social context and shared meanings of a particular group (Geertz, 2000). As such, ethnographic studies yield a vast amount of data—literally hundreds of pages of FIELDNOTES, analytic MEMOS, and/or interview transcripts. In part because of the magnitude of information gathered by ethnographic approaches, the stories social scientists construct from their data may take many different forms. Indeed, although ethnographers share the goal of providing an in-depth look at a group or social phenomenon, they may choose a number of ways to present their findings.

In his book *Tales of the field: On writing ethnography*, Van Maanen (1988) identified three main types of ethnographic tales: realist, confessional, and impressionist. The realist tale aims at presenting an objective view of the group being studied. Thus, the realist tale is akin to the more traditional ethnographic

account wherein readers get a clear, concise description of the individuals or groups under study. They do not get a sense of the author’s reaction to that setting, the impact the author feels he or she has had on the findings, or the reflexive thoughts the author has after returning from “the field.”

The confessional tale, on the other hand, lets the reader in on what the author’s fieldwork experience was like. Although some may argue that the subjective nature of this tale is less scientific and therefore flawed, confessional tales can be a valuable tool in assessing the strength of the research because they allow the reader to get a sense of the author’s approach and biases, as well as the methodological lessons learned while conducting the research. REFLEXIVITY and disclosure of the *process* of research have always been important to feminist qualitative researchers who aim at demystifying the scientific research method in addition to providing meaningful findings.

The impressionist tale is an account of atypical phenomena. The impressionist tale is intended to shock and engage the reader with stark description of unique findings. In this type of tale, the researcher does not offer an *interpretation* of findings but rather brings unusual stories forward to speak for themselves.

Van Maanen’s types of ethnographic tales are not mutually exclusive categories; one may present two or all three types of tales in a narrative account. As Miller (1998) explained in her research on soup kitchens for the homeless, these tales may emerge somewhat unexpectedly as authors reflect on and share their findings. In addition, a primary goal of ethnography and other qualitative methodologies is to *give a voice* to marginalized groups (Ragin, 1994). Giving a voice to individuals whose experiences are not widely recognized *and* providing sound interpretation of research findings may well necessitate the use of various ethnographic tales.

—Desirée Ciambrone

REFERENCES

- Geertz, C. (2000). *The interpretation of cultures: Selected essays*. New York: Basic Books.
- Miller, D. (1998). Writing and retelling multiple ethnographic tales of a soup kitchen for the homeless. *Qualitative Inquiry*, 4, 469–492.
- Ragin, C. C. (1994). *Constructing social research*. Thousand Oaks, CA: Pine Forge Press.
- Van Maanen, J. (1988). *Tales of the field: On writing ethnography*. Chicago: The University of Chicago Press.

ETHNOGRAPHY

Ethnography is the art and science of describing a group or culture. Its strength is description. Ethnography provides the reader with a detailed picture of what's going on, from the perspective of natives of the given culture. It also provides an insight into the processes associated with implementing a program, processes that lead to specific outcomes. One of the primary methodological tools involves fieldwork. Ethnographers immerse themselves in a culture in order to observe and record people's behavior in their natural setting. The time spent observing may range from a few months to a few years; there must be enough time to observe patterns of behavior over time. This is a form of RELIABILITY.

Ethnographers attempt to elicit the insider's or emic perspective of reality while in the field (see EMIC/ETIC DISTINCTION). The ethnographer may or may not agree with the insider's perspective or agree that it conforms to an objective perspective of reality. The aim is to be nonjudgmental, recognizing that there are real consequences for a person's perception of reality, regardless of the scientific merit of that perception. The ethnographer recognizes multiple realities or perceptions of reality based on people's roles in society. For example, the view of the same police officer may be significantly different for an upper-middle-class elderly woman and for a lower-socioeconomic African American male. The ethnographer documents multiple and often conflicting emic or insider perspectives. Most ethnographers build on that knowledge and, using an external or etic perspective, explain the relationship between these emic perspectives of reality.

Ethnography takes more time than other forms of research because it involves fieldwork—time on site with people in their natural environment. Ethnographers spend time in the field in order to place into context the data they collect. Adaptations can be made concerning the amount of time devoted to fieldwork, but there is a direct relationship between time spent in the field and the quality of the data. Ironically, it is often more efficient to be inefficient: Spending more time in the field saves time revisiting poorly understood issues after a study has been completed.

This immersion in the lives of other people allows ethnographers to more accurately interpret people's behavior. This is called cultural interpretation. Clifford Geertz (1975), borrowing Gilbert Riles'

notion of THICK DESCRIPTION, provided a classic in ethnography. It is the example of the wink or blink and how it captures the significance of cultural interpretation in ethnography. The same mechanical movement of an eyelid quickly closing and then opening again might be a wink in a dark and smoky bar or a blink in a dusty furniture repair shop. It is the cultural context that defines how that same mechanical behavior is interpreted (Fetterman, 1998).

FIELDWORK

Fieldwork is one of the most characteristic features of ethnography. Ethnographers typically adopt a naturalistic approach, studying people in their own environment. The most common technique is a stratified judgmental sample, with ethnographers relying on their own judgment to select for study the most appropriate members of the culture in different categories or roles, based on the research question. Questions generally are informal, open, and nonthreatening.

Fieldwork has many phases. The first phase involves entering the field (based on the research proposal). Once introduced to the community, the ethnographer typically begins with a survey period to learn the basics, including language, history, politics, and economics. The postsurvey phase is characterized by continual data collection, theory generation, hypothesis testing, and cross-checking information. In other words, the ethnographer spends a great deal of time guessing about relationships and behavior patterns. Observations and interviews are used to verify, refute, or build on the best guesses about the situation. This approach helps to build a solid foundation of knowledge about the culture or group.

The later phase of fieldwork typically is more narrow and precise in focus. Critical concepts and themes are explored in greater depth (but often less breadth). There is often time pressure to track down missing information, gaps, and contradictions most closely associated with the critical findings or insights associated with the study.

PARTICIPANT OBSERVATION

PARTICIPANT OBSERVATION characterizes most ethnographic work and is crucial to effective fieldwork. Participant observation combines participation in the lives of the people under study with maintenance of a professional distance that allows adequate observation and recording of data. Long-term residence helps the

ethnographer internalize basic beliefs, fears, hopes, and expectations of the group.

CONCEPTS

An ethnographer enters the field with an open mind but not an empty head. Ethnographers have traditional concepts guiding their observation and interpretation. A few of the most significant concepts guiding their work include culture, a holistic perspective, contextualization, emic and etic perspectives, and a nonjudgmental orientation.

Culture

Culture is the broadest ethnographic concept guiding practice. Culture comprises the ideas, beliefs, knowledge, and behavior that characterize a group of people. The concept of culture helps the ethnographer search for a logical, cohesive pattern within a group. Each culture has an identifiable value system that shapes behavior. Discovering the underlying values of a group is like unlocking the human genome. Values constitute a social DNA that helps researchers understand and potentially predict human behavior within a group.

Holistic Perspective

Ethnographers assume a holistic outlook in research, attempting to describe as much as possible about a culture or social group, including its history, religion, politics, and environment. No study can capture an entire culture; however, this orientation pushes the ethnographer to learn more about the depths and breadth of a culture and observe each event within a larger social and economic context. This conceptual lens enables the ethnographer to see interrelationships at the margins and multiple layers of meaning that might be overlooked or ignored in a more focused and narrow approach.

Contextualization

Contextualizing data involves placing observations into a larger perspective. An event or behavior is meaningfully interpreted only when placed in context. A behavior or event viewed in isolation or out of context is easily misinterpreted. Moreover, it is stripped of its multilayered meaning unless situated within its own environment or set of circumstances.

Emic and Etic Perspectives

The emic or insider's perspective is at the heart of most ethnographic research. Ethnographers build their knowledge about a group or culture based on the insiders' views, rather than working from an a priori set of assumptions about how things work. Ethnographers also typically use an etic or external social scientific posture after assembling multiple insider perspectives. The aim is to see how insiders interrelate, communicate, and function as part of a larger system or culture.

Nonjudgmental Orientation

A nonjudgmental orientation helps prevent ethnographers from making inappropriate and unnecessary value judgments about what they observe. A nonjudgmental orientation requires the ethnographer to suspend personal valuation of any given cultural practice. Maintaining a nonjudgmental orientation is similar to suspending disbelief while one watches a movie or play or reads a book—one accepts what may be an obviously illogical or unbelievable set of circumstances to allow the author to unravel a riveting story.

We all have personal beliefs, biases, and individual tastes. The ethnographer guards against the more obvious biases by making them explicit and by trying to view another culture's practices impartially. Ethnocentric behavior—the imposition of one culture's values and standards on another culture, with the assumption that one is superior to the other—is a fatal error in ethnography.

Many other concepts shape or guide an ethnographer in the field, including inter- and intracultural diversity, structure and function, symbol and ritual, micro- or macrolevel study, and operationalism. These concepts are like a theory guiding the ethnographer in practice.

METHODS

The ethnographer is a human instrument. Relying on all its senses, thoughts, and feelings, the human instrument is a most sensitive and perceptive data-gathering tool. The information this tool gathers, however, can be subjective and misleading. Ethnographic methods and techniques help to guide the ethnographer through the wilderness of personal observation and to identify and classify accurately the bewildering variety of events and actions that form a social situation.

Although fieldwork and participant observation characterize much of anthropological work, ethnographers rely on a variety of specific tools, including interviews, SURVEYS, and UNOBTUSIVE MEASURES.

Interviews

Interviewing is the ethnographer's most important data-gathering technique. Interviews explain and put into a larger context what the ethnographer sees and experiences. General interview types include structured, semistructured, informal, and retrospective.

Structured and semistructured interviews serve comparative and representative purposes—comparing responses and putting them in the context of common group beliefs and themes. An informal interview is different from a conversation, but it typically merges with one, forming a mixture of conversation and embedded questions. The questions typically emerge from the conversation. In some cases, they are serendipitous and result from comments made by the participant. The ethnographer uses retrospective interviews to reconstruct the past, asking informants to recall personal historical information.

Types of Questions. There are many types of questions posed in an interview. A survey question or grand tour question is designed to elicit a broad picture of the participant's or native's worldview. Survey questions also help define the boundaries of a study. Once survey questions reveal a category of some significance, specific questions about that category become most useful. Specific questions probe further into an established category of meaning or activity. Whereas survey questions shape and inform a global understanding, specific questions refine and expand that understanding. Structural and attribute questions—subcategories of specific questions—often are used to further clarify the emic perspective.

In addition, ethnographers use both open-ended and closed-ended questions to pursue fieldwork. An open-ended question allows participants to interpret it. Closed-ended questions, with a limited number of answers from which to choose, are useful in trying to clarify views and quantify behavior patterns.

Key Actor and Life History

Some people are more articulate and culturally sensitive than others. These individuals make excellent "key actors," acting as important sources of information for ethnographers. They provide detailed historical data, knowledge about contemporary interpersonal relationships, and a wealth of information about the nuances of everyday life. Although the ethnographer tries to speak with as many people as possible, time is always a factor. Therefore, anthropologists traditionally have relied most heavily on one or two individuals in the group. Key actors often provide ethnographers with rich, detailed autobiographical descriptions or help identify members of the community meriting such depth. These life histories often provide an integrated picture of the culture under study.

Surveys

Surveys are an efficient means of large-scale data collection and an excellent way to tackle questions dealing with representativeness. They are the only realistic way of taking the pulses of hundreds or thousands of people. (See Fowler, 1988, and Fink and Kosecoff, 1988, for guides to conducting surveys.) In addition to surveys, ethnographers use many unobtrusive measures.

Unobtrusive Measures

Unobtrusive measures draw social and cultural inferences from physical evidence. An outcropping, or physical manifestation of an economic or psychological condition, might be graffiti or a burned-out building. Initial inferences are possible without any human interaction. Such cues by themselves, however, can be misleading. Cross-checking and additional data collection are necessary.

Other unobtrusive methods include the study of archival documentation, e-mail, school records, census records, and budgets. Proxemics help differentiate between a close familial relationship and a distant, more businesslike relationship. It also can help us interpret a social situation in which someone's personal space has been entered, with or without permission. Kinesics or body language often signals moods and dispositions (see Birdwhistell, 1970; Hall, 1974).

EQUIPMENT

Notepads, computers, tape recorders, and cameras—all tools of ethnography—are merely extensions of the human instrument, assisting and enhancing memory, capacity, and vision. These useful devices can facilitate the ethnographic mission by capturing the rich detail and flavor of the ethnographic experience and then helping to organize and analyze the data.

The most common tools ethnographers use are pen and paper. With these tools, the fieldworker records notes from interviews during or after each session, sketches the physical layout of an area, traces an organizational chart, and outlines informal social networks. Tape recorders effectively capture long verbatim quotations, essential to good fieldwork, while the ethnographer maintains a natural conversational flow. In addition, TRANSCRIPTION, although time-consuming, is a means of getting closer to the data, becoming more familiar with the patterns of speech, local references, and a person's pauses and tone.

Cameras enable the ethnographer to create a photographic record of specific behaviors. This is a form of FACE VALIDITY. The pictures also can be used as a projective technique; sharing pictures with key actors can elicit useful insights about their behaviors. Photos also can supplement and prompt the ethnographer's memory during the period of data analysis and writing. Digital cameras are now standard in the field. They enable the ethnographer to share pictures immediately—on the camera's screen, attached to e-mail, or stored on a file server for team or community members to view (with permission).

Digital video camcorders are also standard tools in the field. Ethnographers use inexpensive software to produce movies about key events. These represent visual representations or presentations of ethnographic findings. (See Heider, 1976, and Bellman and Jules-Rosette, 1977, for guides concerning ethnographic filmmaking.)

Computers are indispensable in ethnography. Laptops are used in the field to take notes during interviews. Qualitative data analysis software programs such as NUD*IST, ETHNOGRAPH, and ATLAS.ti used to code, chunk, and sort fieldnotes and identify patterns. PowerPoint presentations highlight findings in reports to sponsors and colleagues.

The Internet embodies several tools useful in ethnographic work, including e-mail, search engines, and reference pages, such as phone directories and maps.

In addition, there are many free or inexpensive online survey programs on the Internet. Videoconferencing on the Internet is a way to collect data in the field or to maintain a healthy rapport with key actors while away from the field. These tools save time and money, and they can greatly facilitate communication (see Fetterman, 2002, for additional details).

WRITING

Ethnography requires good writing at every stage. Research proposals, fieldnotes, memoranda, interim reports, final reports, articles, and books, including ethnographies, are the tangible products of ethnographic work. The ethnographer can share these written works with participants to verify their accuracy and with colleagues for review and consideration.

Fieldnotes are the brick and mortar of an ethnographic edifice. These notes consist primarily of data from interviews and daily observation. Ethnographers produce summaries of the research effort during various stages of their fieldwork. The last stage in ethnographic research is writing the final report, article, or book. Ethnographies typically are characterized by thick description and verbatim quotations. Internet publishing represents an alternative to publication in print. Articles can be published and reviewed much quicker on the Internet than they can in traditional media.

ETHICS

Ethics pervade every stage of ethnographic work. Ethnographers must formally or informally seek INFORMED CONSENT to conduct their work. Taking photographs and making audio or video recordings requires the participant's permission (in writing if used for educational or commercial purposes). Ethnographers are candid about their task, explaining what they plan to study and how they plan to study it. They promise confidentiality when requested and use pseudonyms to protect the identity of the individuals with whom they work. Ethnographers need the trust of the people they work with to complete their task. Ethics guide the first and last steps of ethnography, beginning with the selection and definition of the culture and problem, and concluding with the written record of the experience. Ethnographers stand at ethical crossroads throughout the research life cycle, including inception or identification of the problem;

birth, or a funded proposal; childhood, represented by field preparation; adolescence and adulthood, best characterized as fieldwork; and retirement, when it is time to disengage from the field. Each phase poses new challenges as the ethnographer balances methodological concerns with ethical considerations. This fact of ethnographic life sharpens the senses and ultimately refines and enhances the quality of the endeavor. (See Fetterman, 1998, for additional detail; see also the American Anthropological Association's "Code of Ethics" and "Principles of Professional Responsibility" at www.aaanet.org/committees/ethics/ethics.htm.)

—David M. Fetterman

REFERENCES

- Bellman, B. L., & Jules-Rosette, B. (1977). *A paradigm for looking: Cross-cultural research with visual media*. Norwood, NJ: Ablex.
- Birdwhistell, R. L. (1970). *Kinesics and context: Essays on body motion communication*. Philadelphia: University of Pennsylvania Press.
- Fetterman, D. M. (1998). *Ethnography: Step by step* (2nd ed.). Thousand Oaks, CA: Sage.
- Fetterman, D. M. (2002). Web surveys to digital movies: Technological tools of the trade. *Educational Researcher*, 31(6), 29–37.
- Fink, A., & Kosecoff, J. (1988). *How to conduct surveys: A step-by-step guide*. Newbury Park, CA: Sage.
- Fowler, F. J. (1988). *Survey research methods* (Rev. ed.). Newbury Park, CA: Sage.
- Geertz, C. (1975). *The interpretation of cultures*. New York: Basic Books.
- Hall, E. T. (1974). *Handbook for proxemic research*. Washington, DC: Society for the Anthropology of Visual Communication.
- Heider, K. G. (1976). *Ethnographic film*. Austin: University of Texas Press.

ETHNOMETHODOLOGY

Ethnomethodology is the name of an academic field of study as well as a name for what that field studies. The word "ethnomethodology" literally means *popular* or *folk methodology*. The term was coined by Harold Garfinkel (for an account of the origin of the term, see Garfinkel, 1974). After the publication of Garfinkel's (1967) *Studies in Ethnomethodology*, it became an established, if controversial, subfield of sociology. Ethnomethodology has strong connections with

sociology and was recently granted Section status by the American Sociological Association. The connection with sociology runs deep, as ethnomethodology provides a novel approach to the classic themes and problems of sociology; a new volume of Garfinkel's studies (2002) addresses itself to Emile Durkheim's fundamental conception of social facts. Ethnomethodology, however, is not limited to sociology. Together with CONVERSATION ANALYSIS—a spin-off program that investigates the constitutive organization of social interaction—ethnomethodology has made inroads into anthropology, sociolinguistics, computer-supported cooperative work, management studies, and social studies of science, among other fields.

METHODS AS SUBJECT MATTER

Ethnomethodology is not itself a method in any straightforward sense. Just as ethnomusicology is not a branch of music—a particular technique or tradition of music—ethnomethodology is not a kind of methodology. Ethnomethodological studies often make use of ethnographic procedures, and they deploy detailed records of continuous activities "in real time," but in many cases the heaviest burden on the student is to master, and give a procedural account of, specialized methods that constitute the social activities studied.

Ethnomethodology is *sociological* in its fundamental orientation, as it is the study of professional and nonprofessional methods through which social order is produced. Accordingly, ethnomethodologists write procedural descriptions of methods used in a whole range of commonplace and specialized activities. The modern prototype of *method* is the scientific account of experimental or observational procedures. As Sacks (1992, pp. 802–806) noted, scientific methods are communicated by means of vernacular descriptions and prescriptions. Scientific methods can be highly technical, but they are specialized instances of the much broader social phenomenon: instructions that enable members to reproduce a community's practices. In other words, methods and methodological accounts are reflexive features of disciplinary forms of life. Sacks imagined that ethnomethodologists would write methodology reports, but unlike the professional methodologists who write experimental and observational protocols, ethnomethodologists would delve into the rich variety of ordinary methods through which members of a society coordinate and reproduce their activities. These methods include such mundane

“achievements” as crossing the street at a traffic light, coordinating the openings and closings of conversations, bringing a meeting to order, and many other routine actions in daily life.

Ethnomethodological descriptions of lay and professional methods are analogous to methodologies written by scientists and other professionals. Unlike formal methodological accounts, however, they include context-sensitive details of how plans, rules, algorithms, recipes, maps, guidelines, maxims, protocols, proverbs, and other formal instructions are used (as well as disused, misused, or abused) in actual conduct. Such descriptions are not antiformalistic, in the sense of rejecting the relevance of rules and formulas. Instead, ethnomethodological descriptions address the way participants deploy rules, plans, and so forth within singular courses of action. Ethnomethodological studies can require extensive training and PARTICIPANT OBSERVATION in specialized work settings. Often, ethnomethodologists use audio- and videotapes recorded *in situ*, in order to recover the way collective activities are assembled over time through collaborative, often improvisational, sequences of discourse and embodied practice. Typically, this orientation to the phenomenological here and now precludes the use of methods that reduce situated discourse to standard codes and aggregate patterns. Ethnomethodology is not a search for subjective meaning, however. Although informed by existential phenomenology, ethnomethodological descriptions rarely take the form of first-person experiential reflections [an exception is Sudnow's (1978, 2001) elucidation of his own developing competence with playing jazz piano]. More often, such descriptions are written in the third person, without privileging a “private” or individual vantage point. Such descriptions rely upon the intelligibility of common practices to masters of a natural language.

METHODOLOGIES AS SUBJECT MATTER

The subject matter of ethnomethodology includes *methodological* investigations as well as *methodic* practices. In other words, ethnomethodology is the study of *both* methodology (in the sense of systematic reflections about methods and efforts to design them) and methods themselves (practices performed in accordance with one or another methodological design). Such a treatment contrasts with the view that practices are performed unconsciously, requiring professional analysts to account for their systematic organization.

For ethnomethodologists, accounts of practice and practices themselves are inextricable, and it is their intertwining that is of interest.

Garfinkel's program made no exception for social scientific methods, and some of his earliest studies delved into social science research practices such as coding interview responses and analyzing organizational records. These studies demonstrated that *ad hoc* judgments grounded in commonsense knowledge of the social world were crucial for constituting data points and assigning instances to nominal categories. Starting in the 1970s, Garfinkel and some of his students turned their attention to the natural sciences. Like other ethnomethodological studies, investigations of laboratory practices and high-tech workplaces focus on quotidian details, but it is necessary to develop a mastery of the specialized techniques studied in order to demonstrate their ordinarieness. A puzzling feature of these studies is that the sciences already possess their own, highly developed, methodologies. In order to have something to say that is not already part of the sizable professional literatures on methodology, it is necessary for ethnomethodologists to delve into more obscure, and sometimes contentious, features of a practice that are not mentioned in its formal methodological prescriptions and reports. There is some resemblance to social constructionist approaches, but ethnomethodologists do not suggest or imply that the *ad hoc* production of scientific evidence detracts from the objectivity of such evidence. Instead, for ethnomethodologists, the tacit features of scientific production are unavoidable; they are a source of objectivity as much as a barrier to it.

—Michael Lynch

REFERENCES

- Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice Hall.
- Garfinkel, H. (1974). On the origins of the term “ethnomethodology.” In R. Turner (Ed.), *Ethnomethodology* (pp. 15–18). Harmondsworth, UK: Penguin.
- Garfinkel, H. (2002). *Ethnomethodology's program: Working out Durkheim's aphorism*. Blue Ridge Summit, PA: Rowman and Littlefield.
- Sacks, H. (1992). *Lectures on conversation* (Vol. 1). Oxford, UK: Blackwell.
- Sudnow, D. (1978). *Ways of the hand*. Cambridge, MA: Harvard University Press.
- Sudnow, D. (2001). *Ways of the hand* (Rev. ed.). Cambridge, MA: MIT Press.

ETHNOSTATISTICS

Ethnostatistics uses concepts from ETHNOMETHODOLOGY to study sensemaking practices that social scientists employ in the production, interpretation, and display of statistics created in social research. Ethnostatistics focuses on quantitative reasoning and language use in social contexts to understand how professionals produce numerals that create a sense of shared social reality.

Ethnostatistics contrasts with STATISTICS, which involves techniques and rule-governed procedures for counting, aggregating, and estimating data. The field of statistics does not study practical knowledge needed to produce statistics or how situations influence the behavior of the scientist. Formal statistical procedures and rules are thus incomplete guides to producing statistics. Ethnostatistics is concerned with describing and understanding the technical knowledge and procedures needed to actually do statistics.

Ethnostatistics has three levels. *First-level ethnostatistics* uses ethnography to produce THICK DESCRIPTIONS of the social production of statistics. For example, Gubrium and Buckholdt (1979) studied the professional production of quantitative measures of hospital program effectiveness and demonstrated that technical rules are insufficient to explain counting practices and statistics. Informal criteria, such as patient motivation to comply with program goals, were used in addition to technical criteria by professional staff to determine which actions to count.

Second-level ethnostatistics studies statistics at work using COMPUTER SIMULATION of the implications for the interpretation of statistics that are associated with violations of various technical assumptions. One study tested the validity of the common social science assumption that ordinal data can be transformed into interval data for statistical analysis purposes by assigning successive numerals (e.g., 1–7) to data categories. This practice assumes that data transformation does not affect the interpretation of results and findings. Ethnostatistical research shows that this transformation can produce measured values and statistics that diverge “substantially from true measures” (Gephart, 1988, p. 38). Hypothesis testing and PATH ANALYSIS interpretations of data also were affected. Second-level ethnostatistics thus shows that common measurement practices may distort results.

Third-level ethnostatistics uses RHETORIC and textual analysis methods to examine persuasive properties of statistical displays in documents and texts. One study of how quantitative methodology papers create meaning and interpret numerals found that meaning is created by the adjectives that are used to describe and interpret numerals and by contexts where statistical descriptions are embedded. The absolute value of a numeral did not govern the interpretation made of the numeral, and tests of statistical significance played a minor role. Indeed, the correlation an author termed “substantial” in a quantitative paper often was within .1 of a correlation claimed to be “relatively small” at another point in the article.

Ethnostatistics integrates the quantitative and qualitative aspects of science by examining how meaning is actually created for numbers. This re-humanizes quantitative research by relating it to qualitative foundations of social life (Gephart, 1988).

—Robert P. Gephart, Jr.

REFERENCES

- Gephart, R. P. (1986). Deconstructing the defence for quantification in social science: A content analysis of journal articles on the parametric strategy. *Qualitative Sociology*, 9, 126–144.
- Gephart, R. P. (1988). *Ethnostatistics: Qualitative foundations for quantitative research*. Thousand Oaks, CA.: Sage.
- Gephart, R. P. (1997). Hazardous measures: An interpretive textual analysis of quantitative sensemaking during crises. *Journal of Organizational Behavior*, 18, 583–622.
- Gubrium, J. F., & Buckholdt, D. R. (1979). Production of hard data in human service organizations. *Pacific Sociological Review*, 22, 115–136.

ETHOGENICS

The word “ethogenics” was coined to refer to a new paradigm for social psychological research that developed in Oxford, England, in the 1970s (Harré & Secord, 1972) in conscious reaction to the importation of methods and theories from mainstream American psychology. Ethogenics was influenced by ETHNOMETHODOLOGY (Garfinkel, 1967), by the microsociology of Goffman (1959), and by social and cultural anthropology.

The movement rejected universalistic presuppositions that had led to levels of abstraction that eliminated

concrete social phenomena. The treatment of persons as passive sites for stimulus/response patterns was rejected in favor of treating persons as active agents engaged with others in carrying out projects according to local rules and conventions. Fallacious uses of statistical methods were highlighted, in particular the familiar but still common fallacy of drawing conclusions about individual propensities from statistical distributions.

Emphasis was placed on the analysis of actual episodes of social interaction as unfolding sequential structures of meanings, ordered in accordance with local rules, conventions, and customs of correct conduct. Studies of the dynamics of social action—for example, the *making* of friends—displaced studies of the conditions under which static states, such as *friendship*, were brought about. The focus on local customs and conventions led to an interest in social anthropology as an essential component of a new paradigm of social psychology.

The methodological characteristic of research within the new paradigm required the use of the act/action distinction to identify social phenomena as meanings and how those meanings were understood by the participants. Observation of real-life episodes replaced EXPERIMENTS. Analysis of participants' justificatory and interpretative accounts was used to identify the projects to the accomplishment of which a social interaction was directed, as well as the rules and conventions in accordance with which it was managed.

The upshot of research within the new paradigm was a catalog of situation-specific meanings and sets of context-sensitive rules that explained the pattern of the evolving social episode, viewing it as an actual sequence of meaningful social actions. This led to an alliance with the newly emergent field of discursive psychology, which was directed to a similar end product. The concept of "rule" stood in for a wide variety of normative constraints that could be seen to be effective in shaping the flow of action.

Ethogenics placed great emphasis on the construction of descriptive and explanatory models. The most powerful of these was the dramaturgical model. Social episodes can be looked at as if they were performances of stage plays, with scene, actor, and action mutually influencing one another. This basic model was elaborated by exploring analogies that compared certain kinds of social episodes to games and ceremonies.

In later applications of this point of view in developmental social psychology, alliances grew up

with (SOCIAL) CONSTRUCTIONISM and with cultural psychology (Cole, 1996).

—Rom Harré

REFERENCES

- Cole, M. (1996). *Cultural psychology: A once and future discipline*. Cambridge, MA: Belknap Press of Harvard University Press.
- Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice Hall.
- Goffman, E. (1959). *The presentation of self in everyday life*. New York: Doubleday.
- Harré, R., & Secord, P. F. (1972). *The explanation of social behaviour*. Oxford, UK: Blackwell.

ETHOLOGY

Ethology is a prime example of a discipline whose models and methods were developed in nonhuman animal research and then applied to research on humans. Firmly grounded in the evolutionary thinking of Charles R. Darwin, ethology perhaps made its most valuable contribution to theoretical refinement through its commitment to understanding behavior on multiple levels: (a) proximate causation (which may be physiological or environmental in origin), (b) ultimate causation (the adaptive function of the behavior in enhancing survival and/or reproductive fitness), (c) phylogeny (the behavior's pattern in related species), and (d) ontogeny (the developmental course of the behavior) (Tinbergen, 1963). This general framework led animal researchers to leave their controlled laboratories and seek out natural habitats where species of interest could be observed in the environmental niches to which they had adapted. For example, Konrad Lorenz discovered the phenomenon of imprinting in geese by observing goslings as they hatched and began following either their mothers or Lorenz himself soon after emerging from their eggs (Lorenz, 1981). It was only after completing the descriptive phase of this observational work that Lorenz began to experiment with imprinting, for example, by introducing time since hatching as a variable, still in the natural setting of a lakeshore in Germany. Thus, ethology is often defined as the study of animal behavior in the animal's natural environment.

Describing behavior in its natural setting does not imply abandoning scientific ideals of rigorous research.

An ethologist will spend countless hours cataloging behaviors of the subject species, thus creating an “ethogram” for that species. Specific questions typically are asked: Which behaviors seem to occur together? Under what conditions in the environment? Do the behaviors seem directed toward a goal? What does the animal do if the goal is not achieved? Do the behaviors make up a fixed action pattern? How often do the behaviors appear? Are there individual differences, perhaps related to sex, age, or social status? Are there differences related to time of day, season, temperature, or other environmental signals? Throughout the process, the ethologist will ideally return to Tinbergen’s four levels of analysis cited earlier, for theoretical refinement. Where appropriate, the ethologist also will utilize experimental methods to answer questions. For example, when a specific environmental stimulus is hypothesized to be a “releaser” of a behavior, the ethologist may carry out a deprivation experiment, eliminating the stimulus, to see how the behavior of interest is modified. By manipulating the timing of releasers, ethologists have identified sensitive periods (formerly called critical periods) in the development of such behaviors as imprinting in geese and socialization in dogs. Thus, ethological research may be described as moving back and forth in a systematic way between qualitative and quantitative approaches to studying behavior.

During the 1960s and 1970s, some scientists trained in classical ethology moved toward human research, as seen in the work of Irenaeus Eibl-Eibesfeldt (1989), who studied with Lorenz and documented human emotional communication around the world. It was also at this time that John Bowlby and Mary Ainsworth, influenced by the evolutionary approach of Darwin and the primate methodology of Harry Harlow, captured the attention of psychology with their work on mother-infant attachment. Many of those who admired the methodology, however, failed to appreciate Bowlby’s evolutionary viewpoint, particularly in proposing infant survival as the ultimate goal of attachment.

Ethology takes a meticulous approach to OBSERVATIONAL RESEARCH. Several sources outline in great detail ethological standards for conceptualizing questions, SAMPLING subjects, recording time and space, recording events, measuring environmental stimuli, and evaluating observers’ reliability (see, e.g., Blurton Jones, 1972; LaFreniere & Charlesworth,

1983; and McGrew, 1972). For example, if one is interested in how often preschoolers engage in aggressive interactions, one would first define aggression and list observable behaviors that fit the definition (push, bite, kick, etc.) Then, in preliminary stages of a study, one might note easily observed acts of aggression (*ad libitum* sampling); or one might note how many times those behaviors are observed every minute (event sampling); or one might note whether or not an aggressive behavior is observed every 15 seconds (one-zero sampling); or one might watch a certain child and note whether or not that child exhibits an aggressive behavior in a given time period (focal sampling); or one might note what comes before and after aggressive behaviors (sequence sampling). By utilizing these methods in an ethological study of nursery schoolers, McGrew (1972) demonstrated (among other findings) that aggression is related to winning/losing property fights and that teachers’ ratings correlate highly with scientifically observed behavior.

At times, the need for more fine-grained analysis will necessitate the use of film or video analyzed in slow motion, or the use of sound recordings analyzed in the laboratory, or the use of chemical analysis to break down olfactory stimuli. In human research, facial muscle movements have been broken down into single units, and posture and movement quality are now being explored through computer modeling in the work of Karl Grammer, Bernhard Fink, and LeeAnn Renninger (2002).

In assessing the importance of ethology, Tinbergen (1963) commented, “It has been said that, in its haste to step into the twentieth century and to become a respectable science, Psychology skipped the preliminary descriptive stage that other natural sciences had gone through, and so was soon losing touch with the natural phenomena” (p. 411). Ethology, which has the advantage of providing both descriptive and experimental research methods, may have untapped heuristic value for the social sciences in general. In addition, ethology’s underpinnings in evolutionary theory offer an inclusive approach that provides a way out of the fruitless nature/nurture dichotomy.

In recognition for their work in ethology, Konrad Lorenz, Nikolaas Tinbergen, and Karl von Frisch were awarded the Nobel Prize in Physiology or Medicine in 1973.

—Carol Cronin Weisfeld

REFERENCES

- Blurton Jones, N. (Ed.). (1972). *Ethological studies of child behaviour*. London: Cambridge University Press.
- Eibl-Eibesfeldt, I. (1989). *Human ethology*. Hawthorne, NY: Aldine de Gruyter.
- Grammer, K., Fink, B., & Renninger, L. (2002). Dynamic systems and inferential information processing in human communication. *Neuroendocrinology Letters*, 23(Suppl. 4), 15–22.
- LaFreniere, P., & Charlesworth, W. R. (1983). Dominance, attention, and affiliation in a preschool group: A nine-month longitudinal study. *Ethology and Sociobiology*, 4(2), 55–67.
- Lorenz, K. (1981). *The foundations of ethology*. New York: Simon and Schuster.
- McGrew, W. (1972). *An ethological study of children's behavior*. New York: Academic Press.
- Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, 20, 410–433.

EVALUATION APPREHENSION

Evaluation apprehension is a form of reactive effect that has been shown to occur in EXPERIMENTS, whereby the experimental subject becomes worried that he or she is being tested by the experimenter and that he or she may perform poorly. It is sometimes recommended that the COVER STORY for an experiment should make it clear that the experiment has not been designed to evaluate the performance of participants.

—Alan Bryman

EVALUATION RESEARCH

Evaluation research is systematic, data-based inquiry to determine the merit or worth of a program, product, organization, intervention, or change effort. Evaluation research applies social science and related inquiry methods for the systematic collection of information about the activities, characteristics, and outcomes of change efforts to inform judgments about goal attainment, improve program effectiveness, identify costs and benefits, and/or inform future decisions.

Billions of dollars are spent to fight problems of poverty, disease, ignorance, joblessness, mental anguish, crime, hunger, and inequality, to name but a few of the many efforts at improving the human

condition. How are programs that combat these societal ills to be judged? How does one distinguish effective from ineffective programs? How can information be gathered and reported in ways that increase program effectiveness and enhance decision making? These questions are the domain of evaluation research.

EVALUATION AND THE EXPERIMENTING SOCIETY

Edward Suchman (1967) began his seminal text on evaluation research with the observation that “one of the most appealing ideas of our century is the notion that science can be put to work to provide solutions to social problems” (p. 1). Visionaries such as Donald T. Campbell (1991) conceptualized evaluation as the centerpiece of a new kind of society: *the experimenting society*. He gave voice to this vision in his 1971 address to the American Psychological Association (1991):

The experimenting society will be one which will vigorously try out proposed solutions to recurrent problems, which will make hard-headed and multidimensional evaluations of the outcomes, and which will move on to other alternatives when evaluation shows one reform to have been ineffective or harmful. We do not have such a society today. (p. 223)

HISTORICAL CONTEXT

Evaluation in the United States traces its formal professional roots to the federal projects spawned by the Great Society legislation of the 1960s, which were aimed at nothing less than the elimination of poverty. The creation of large-scale federal health programs, including community mental health centers, was coupled with a mandate for evaluation, often at a level of 1% to 3% of program budgets. Other major programs were created in housing, employment, services integration, community planning, urban renewal, and welfare. The rapid growth of social action programs created a demand for systematic empirical evaluation of the effectiveness of government programs. Passage of the U.S. Elementary and Secondary Education Act in 1965 contributed greatly to more comprehensive approaches to evaluation. The massive influx of federal money aimed at desegregation, innovation, compensatory education, greater equality of

opportunity, teacher training, and higher student achievement was accompanied by calls for evaluation data to assess the effects on the nation's children of various programs.

Program evaluation as a distinct field of professional practice was born of two lessons from this period of large-scale social experimentation and government intervention: first, the realization that there is not enough money to do all the things that need doing; and second, even if there were enough money, it takes more than money to solve complex human and social problems. *Because not everything can be done, there must be a basis for deciding which things are worth doing.* Evaluation research was charged with informing decision making by bringing credible data to bear on judgments about effectiveness.

Early visions for evaluation, then, focused on evaluation's expected role in guiding funding decisions and differentiating the wheat from the chaff in federal programs. As evaluations were implemented, a new role emerged: *helping improve programs as they were implemented.* Evaluators were called on not only to offer final judgments about the overall effectiveness of programs but also to gather process data and provide feedback to help solve programming problems along the way.

MAJOR DEVELOPMENTS IN EVALUATION RESEARCH

1. Emergence of Evaluation as Both Discipline and Profession

Evaluation research began as the application of methods to particular problems within established disciplines such as economics, psychology, and education. As evaluation researchers began to come together to share experiences and perspectives, evaluation emerged as a distinct field of professional practice. Formal professional associations with annual meetings and professional journals were formed in the 1970s and 1980s, among them the American Evaluation Association, the Canadian Evaluation Society, and the Australasian Evaluation Society.

Evaluation also became a scholarly area of inquiry specializing in how merit, worth, value, and effectiveness are determined, thereby emerging out of the shadows of other disciplines to become a discipline in its own right, or even, as philosopher and

evaluator Michael Scriven (2001) has suggested, a *transdiscipline*.

[E]valuation is one of the elite group of trans-disciplines, a term used in this context to refer to disciplines that are most notable for their service to other disciplines, although having their own autonomous status as well. These transdisciplines range from little but important ones like measurement, up through major ones like probability and statistics, key tools for most of the quantitative disciplines, to the all-encompassing ones—logic and evaluation. (p. 304)

2. Standards of Excellence for Evaluation

The professionalization of evaluation led to articulation of standards. Before the profession identified and adopted its own standards, criteria for judging evaluations could scarcely be differentiated from criteria for judging research in the traditional social sciences, namely, technical quality and methodological rigor. The most comprehensive effort at developing standards was hammered out over 5 years by a 17-member committee appointed by 12 professional organizations, with input from hundreds of practicing evaluation professionals. In 1981, the standards were published, then revised in 1994 (Joint Committee on Standards for Educational Evaluation, 1994). The standards dramatically reflected the ways in which the practice of evaluation had matured, calling for evaluations to be judged by four sets of criteria: utility, feasibility, propriety, and accuracy.

3. Increasing Attention to Evaluation Utilization

By the end of the 1960s, it was clear that evaluations of Great Society programs were largely ignored or politicized. Carol Weiss (1972) was one of the first to call attention to underutilization as one of the foremost problems in evaluation research, in a seminal review that concluded that "evaluation results have not exerted significant influence on program decisions" (pp. 10–11).

Nor was the challenge only one of increasing use. Evaluators became increasingly concerned about the misuse of findings. Marvin Alkin (1990), an early theorist of user-oriented evaluation, emphasized that

evaluators must attend to *appropriate* use, not just amount of use.

Thus, evaluation faced a dual challenge: enhancing appropriate uses while also eliminating improper uses. In response to these concerns, an approach called UTILIZATION-FOCUSED EVALUATION emerged (Patton, 1997), with emphasis on conducting evaluations to achieve intended use by intended users. Utilization-focused evaluation involves designing an evaluation with careful consideration of how everything that is done, *from beginning to end*, will affect use. A psychology of use undergirds utilization-focused evaluation: Intended users are more likely to use evaluations if they understand and feel ownership of the evaluation process and findings; they are more likely to understand and feel ownership if they've been actively involved; and by actively involving primary intended users, the evaluator is training users in use, preparing the groundwork for use, and reinforcing the intended utility of the evaluation at every step along the way.

Utilization-focused evaluation also brought attention to *process use* (Patton, 1997), in contrast to findings use. Process use refers to changes in thinking and behavior, as well as program or organizational changes in procedures and culture, that occur among those involved in evaluation as a result of the learning that occurs during the evaluation process. Process use is represented by program staff saying after an evaluation, "The impact on our program came not so much from the findings but from going through the thinking process that the evaluation required."

4. Methodological Diversity Valued

Early in the development of evaluation research, an intense debate raged about the relative merits of quantitative/experimental methods versus qualitative/naturalistic methods. That debate has run out of intellectual steam (Patton, 1997). A consensus has emerged in the profession that evaluators need to know and use a variety of methods in order to be responsive to the nuances of particular evaluation questions and the idiosyncrasies of specific stakeholder needs. The focus in designing evaluation research is now the appropriateness of methods for specific purposes and intended uses, *not* adherence to some absolute orthodoxy that quantitative or qualitative methods should be inherently preferred. The Joint Committee Standards (1994) give equal weight to both kinds of methods. The field has come to recognize

that, where possible, using multiple methods—both quantitative and qualitative—can be valuable because each has strengths and one approach can often overcome weaknesses of another.

5. Many Different Types of Evaluation and New Evaluator Roles

Parallel to increased methodological diversity, the practice of evaluation research has become diversified with many new approaches, models, and evaluator roles. Initially, evaluations simply assessed a program's goal attainment, making an overall judgment about merit or worth called a *summative evaluation* or *judgment-oriented evaluation*. Summative evaluations measure outcomes and impacts, and the causal linkages between program activities and outcomes that constitute a program model or theory. Gradually, more attention came to be placed on using evaluation to improve programs by getting participant feedback and identifying strengths and weaknesses, what is often called *formative evaluation* or improvement-oriented evaluation. Formative evaluations examine implementation, program processes, and program adaptations to changing clients or conditions. Sometimes formative evaluations serve to get programs ready for summative evaluation, but increasing use of improvement-oriented evaluation has become part of ongoing efforts for continuous quality improvement.

Both judgment-oriented and improvement-oriented evaluations involve the *instrumental* use of results, meaning that a decision or action follows from the evaluation. Conceptual use of findings, on the other hand, increases knowledge so that evaluation findings can influence thinking about issues, options, or policy alternatives. The knowledge generated can be as specific as clarifying a program's model, testing theory, distinguishing types of interventions, figuring out how to measure outcomes, generating lessons learned, and/or elaborating policy options. In recent years, knowledge-generating evaluations have come to be highly valued.

Another shift involved external versus internal evaluation. With the early focus on summative evaluation, most evaluations were commissioned externally to ensure independence and credibility. With greater emphasis on continuous quality improvement, internal evaluation began to flourish.

New approaches to evaluation have emerged to solve particular problems or accomplish specific

objectives. *Goal-free evaluation* offers an alternative to goals-based evaluation as a way of finding out whether clients' primary needs were met and of capturing side effects and unintended impacts. *Responsive evaluation* is an approach to capturing and taking into account different and possibly conflicting stakeholder perspectives. *Empowerment evaluation* is a way to help participants tell their own stories and increase their capacity to engage in evaluation processes. *Inclusive evaluation* emphasizes the need for special efforts to include the perspectives of the disadvantaged and less privileged. *Theory-driven evaluation* formally tests social science theories in real program settings. *Realist evaluation* uses the philosophy of REALISM to build an evaluation framework. *Democratic evaluation* focuses on using evaluation to support dialogue and deliberation to enhance social justice. In *developmental evaluation*, the evaluator is part of a team whose members collaborate to conceptualize, design, and test new approaches in an ongoing process of continuous improvement.

These approaches offer but a sampling of the rich and varied menu of options that have emerged in evaluation research.

6. New Skills and Competencies Needed

New approaches to evaluation research have broadened the skills needed to be an effective evaluator. Evaluation research has developed beyond the simple application of social science research methods to assess a program's goal attainment. To ensure utility, evaluators need skills in communication, negotiation, conflict resolution, and group facilitation. To meet the standards of feasibility, evaluators need to be politically sophisticated, pragmatic, and creative, as well as able to manage evaluation processes so as to produce results on time and within budget. To meet standards of propriety, evaluators need a keen sense of ethics and sensitivity to diverse stakeholder groups. To meet standards of accuracy, evaluators need to be able to employ a variety of methods and work across disciplines. These new demands are changing the ways that evaluation researchers are trained.

7. Technology and Evaluation Research

The dominant role of technology in modern society has affected evaluation research as it has affected other endeavors. The Internet has opened up new communication channels as evaluators around the

world are linked together through networks and listservs. Advances in software facilitate data analysis, including especially major developments in software for analyzing qualitative data (Patton, 2002). Advances in computer graphics and word processing software have changed how evaluation reports are written and disseminated. Placing major evaluation reports on the World Wide Web has increased access to evaluation findings, facilitated the dissemination of evidence-based practices, and enabled more work in synthesizing lessons learned across different evaluations, a form of generating knowledge. New techniques for data gathering using the Internet (e.g., Web-based surveys) are changing how data are gathered. Videotaping and photography are used in both data gathering and reporting. Management information systems are assisting programs in gathering routine monitoring data that can be aggregated for program evaluation purposes. In these and many other ways, evaluation research is influenced by technological developments.

8. Worldwide Demand for Evaluation

The first international evaluation conference in Vancouver in 1995 was planned by the three major evaluation societies active at that time: the Canadian, the Australasian, and the American. With more than 1,500 participants from 61 countries, that conference remains one of the defining moments of evaluation's history. By early in the 21st century, more than 40 national evaluation associations had been formed, in addition to the European Evaluation Society, the African Evaluation Association, and an International Evaluation Association made up of national and regional associations. International agencies have also begun using evaluation to assess the full range of development efforts under way in developing countries. Global interactions are defining the future of evaluation research, infusing the profession with new energy as evaluators around the world learn from each other and grapple with the ways in which evaluation research has to be adapted to fit different cultural, societal, and political contexts.

—Michael Quinn Patton

REFERENCES

- Alkin, M. (1990). *Debates on evaluation*. Newbury Park, CA: Sage.
- Campbell, D. T. (1991). Methods for the experimenting society. *Evaluation Practice*, 12, 223–260.

- Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards*. Thousand Oaks, CA: Sage.
- Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.
- Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Rossi, P. H., Freeman, H. E., & Lipsey, M. (1998). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.
- Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.
- Scriven, M. (2001). Evaluation: Future tense. *The American Journal of Evaluation*, 22, 301–307.
- Shadish, W. R., Jr., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.
- Shadish, W. R., Jr., Newman, D. L., Scheirer, M. A., & Wye, C. (1995). *Guiding principles for evaluators* (New Directions for Program Evaluation, No. 66). San Francisco: Jossey-Bass.
- Suchman, E. A. (1967). *Evaluative research: Principles and practice in public service and social action programs*. New York: Russell Sage.
- Weiss, C. H. (1972). A treeful of owls. In C. H. Weiss (Ed.), *Evaluating action programs* (pp. 3–27). Boston: Allyn & Bacon.
- Weiss, C. H., & Bucuvalas, M. (1980). Truth test and utility test: Decision makers' frame of reference for social science research. *American Sociological Review*, 45, 302–313.

EVENT COUNT MODELS

In many research contexts, the dependent variable is measured as a count. A count variable is a non-negative integer that represents the number of observed events (or people, items, etc.). The objective is to explain the variation in observed counts by incorporating predictor variables. ORDINARY LEAST SQUARES REGRESSION analysis, which assumes a CONTINUOUS dependent VARIABLE, is not well suited for count data.

Count models are best thought of within the context of the GENERALIZED LINEAR MODEL. Any PROBABILITY distribution that allows only nonnegative integers can be used to describe the dependent variable. A careful study of count models can serve as an excellent bridge into the literature on the generalized linear model.

Suppose there is a vector of m predictor variables for each observation i : $X_i = (x_{1i}, x_{2i}, \dots, x_{mi})$. Suppose that estimated coefficients $b = (b_1, b_2, \dots, b_m)$, can be used to predict the expected value of the dependent

variable, y_i . One's first instinct might be to use the linear model, such as

$$E(y_i|X_i) = X_i b = \sum_{j=1}^m b_j x_{ji}.$$

The shortcoming of this approach is that the expected value of y_i might be negative, thus violating the basic ASSUMPTION that all observed counts are 0 or greater. To circumvent that problem, a mathematical TRANSFORMATION typically is used. Most commonly, one finds the exponential transformation:

$$E(y_i|X_i) = \exp(X_i b) = e^{X_i b} = e^{\sum_j b_j x_{ji}}.$$

The MAXIMUM LIKELIHOOD ESTIMATION approach is used to derive the estimates of b .

The maximum likelihood calculations require that some specific statistical distribution for y_i must be assumed. Until about 1990, by far the most commonly used distribution was the POISSON DISTRIBUTION. Given input variables X_i and parameter estimates $\hat{b}$, the probability of observing exactly y_i events is

$$\Pr(y_i|X_i) = \frac{\exp(-e^{X_i b})(e^{X_i b})^{y_i}}{y_i!}.$$

The estimates of the parameters b are chosen so as to make the sample of observations most likely.

The Poisson distribution is relatively simple and workable. If $E(y_i|X_i)$ is small, then the Poisson distribution is sharply skewed, with a very small modal value and a long tail to the right. As $E(y_i|X_i)$ increases, the shape changes and becomes rather normal in appearance. (The similarity between the Poisson distribution for large values and the normal distribution sometimes is used to justify the use of ordinary least squares regression with count data.) For a survey of the Poisson model and social science applications, consult King (1988).

The Poisson distribution has an unsuitable property: Its VARIANCE is equal to its mean. When it is fitted to observed data, one often finds the problem of overdispersion; the variance of the observed data is greater than expected according to the Poisson distribution.

Many scholars propose replacing the Poisson distribution with the NEGATIVE BINOMIAL (NB) DISTRIBUTION. The NB model has the same mean as the Poisson, but it has greater variance. It has another useful property. We can write the original model in a way that allows individual HETEROGENEITY in the

form of an additive DISTURBANCE (or error) TERM. Let the (unobserved) individual variation be represented by u_i :

$$\begin{aligned} E(y_i|X_i) &= \exp(X_i b + u_i) = e^{X_i b + u_i} \\ &= e^{\sum_j b_j x_{ji} + u_i} = e^{\sum_j b_j x_{ji}} \times e^{u_i}. \end{aligned}$$

If the variance of u_i happens to be 0, then this degenerates to a Poisson model (because $e^0 = 1$). However, if we posit that u_i has a GAMMA distribution, then the observed distribution of y_i would be the negative binomial. It is typical to assume that $E(u_i) = 0$, so that “on average” the heterogeneity observed among cases has no effect. That is to say, the expected value of a Poisson or negative binomial model is the same. However, when the variance of $u_i > 0$, the dispersion of the negative binomial distribution is greater. Long’s (1997) textbook offers an excellent treatment of this issue.

Many extensions of the count model framework are available or are being pioneered in the advanced literature of statistics. The approach can be extended to situations in which one observes a greater than expected number of zeroes in the counts. An encyclopedic treatment of count models was presented by Cameron and Trivedi (1998).

—Paul E. Johnson

REFERENCES

- Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge, UK: Cambridge University Press.
- King, G. (1988). Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential Poisson regression model. *American Journal of Political Science*, 32, 838–863.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

EVENT HISTORY ANALYSIS

Event history analysis is a technique that allows researchers to address not only whether an event occurs but also when it occurs. An event is a change from one state to another, and the dependent variable is the time until the event occurs. Event history analysis is ideally suited to the study of longitudinal change and can be

thought of as extending LOGIT/PROBIT ANALYSIS and EVENT COUNT MODELS to take into consideration the timing of the event(s).

Two common complications arise in longitudinal data analysis that motivate the use of event history analysis. First, censored observations exist in the data when information about the duration (the amount of time an observation spends in a particular state) is incomplete. This may occur, for example, because the observation did not experience the event of interest prior to the end of the study or because the observation is lost in follow-up, perhaps because the subject moved and could not be located. Second, time-varying covariates (or independent variables) have values that change over time. For example, in a study of the timing of challenger entry in a congressional election, the amount of money raised by an incumbent legislator could be a time-varying explanatory variable across the election cycle. Event history techniques can readily incorporate censored observations and time-varying explanatory variables. The inclusion of time-varying covariates in event history analysis can lead to novel information regarding how the risk of an event occurrence changes in relation to changes in the value of that covariate.

HISTORICAL DEVELOPMENT

Event history analysis is also referred to as *duration*, *survival*, or *reliability analysis*, depending on the substantive origins of the discussion (medicine and engineering for the latter two terms, respectively). Early applications involved life table analysis by Kaplan and Meier (1958), but the historical roots can be traced back even further to the late 1600s (Hald, 1990). There was an increased use of the technique during World War II because of concerns over the expected reliability of military equipment. The path-breaking work of D. R. Cox (1972) is credited with another period of expansion in the use of event history analysis as a result of his development of semiparametric techniques. His work is likely to be heralded as one of the top statistical achievements in the 20th century. Applications in medicine and the social sciences have increased greatly as a result of the less restrictive semiparametric Cox regression model and its various extensions, which are built upon the mathematics of counting processes (Therneau & Grambsch, 2000).

PARAMETRIC AND SEMIPARAMETRIC MODELS

Both parametric and semiparametric models are available in event history analysis. Analysts studying mechanical systems typically use parametric models, which assume that the time until the event of interest follows a specific distribution, such as the exponential. Studies of human behavior and biology typically use the less restrictive semiparametric Cox model, which leaves the particular distributional form of the duration times unspecified. Blossfeld and Rohwer (2002, pp. 180, 263) argued that social science theory rarely provides the justification for a specific parametric distribution and instead advocated for use of the Cox model.

A key concept in the estimation and interpretation of an event history model is the HAZARD RATE. The hazard rate gives the rate at which observations fail (or durations end) by time t given that the observation has survived through time $t - 1$. In other words, it can be interpreted as the probability that an event will occur for a particular observation at a particular time. The risk set refers to those observations that are still "at risk" of experiencing the event of interest. Once the observation experiences the event at time t , the observation exits the risk set and is no longer part of the data set being analyzed at $t + 1$. The hazard rate has substantive appeal in that the event of interest is conditional on its history. For example, given that a war has lasted t periods, what is the likelihood that it will end in the subsequent period? The hazard rate for the Cox model may be written as

$$h(t|X_i) = h_0(t)e^{X_i\beta},$$

where $h_0(t)$ is an (unspecified) baseline hazard function and X_i are covariates for observation i .

MODEL ASSUMPTIONS AND MODEL FITTING

The major assumption to be checked when fitting event history models is the proportional hazards assumption. Most event history models, including the Cox model, assume that the hazard functions of any two individuals with different values on one or more covariates differ only by a factor of proportionality. Put differently, the baseline hazard rate varies with time but not across individuals, so that the ratio of the hazards

for individuals i and j are independent of t and are constant for all t :

$$\frac{h_i(t)}{h_j(t)} = e^{\beta(X_j - X_i)}.$$

Estimation of Cox's model when hazards do not satisfy the proportionality assumption can result in BIASED and inefficient estimates of *all* parameters, not simply those for the covariate(s) in question. The proportional hazards assumption should be checked with Harrell's ρ for individual covariates and with Grambsch and Therneau's global test for nonproportionality (Box-Steffensmeier & Zorn, 2001; Therneau & Grambsch, 2000). If evidence of nonproportionality is found (and, in most social science research, proportionality is more the exception than the rule), then the potentially nonproportional covariates should be interacted with $\ln(\text{time})$ or other appropriate transformations of time. Such interactions allow each interacted covariate's effect on the hazard of conflict to vary monotonically with the duration of the event being studied. Relaxing this assumption allows scholars to test whether the effects of covariates change over time and permits a more nuanced understanding of the phenomenon being studied. Moreover, nonproportionality tests, and the residuals upon which they are based, are increasingly easy to obtain in commonly used software packages for analyzing duration data.

If parametric models are estimated, the assumption of the chosen parametric distribution needs to be tested, and the proportional hazard assumption may still need to be assessed (for example, the Weibull model also assumes proportional hazards). The generalized gamma distribution is an encompassing model for several commonly used parametric distributions and thus may serve to help adjudicate among competing nested models. Because the parametric models are estimated by maximum likelihood and the properties of these estimators are well known, the standard battery of goodness-of-fit indices and statistics are directly applicable to the parametric modeling framework, for example, use of the LIKELIHOOD RATIO test or the Akaike information criterion (AIC). However, the principal advantage of the Cox model is not having to make assumptions about the nature and shape of the baseline hazard rate, and thus the Cox model should be the first choice among modeling strategies for social scientists (Box-Steffensmeier & Jones, 1997).

Another underlying assumption of almost all event history models is that all observations eventually will experience the event of interest. This assumption can be relaxed by estimating a split-population model. (These models are also known as cure models in biostatistics, a name based on the idea that part of the population is cured.) As examples, in studies of the timing of campaign contributions, split-population models do not assume that every political action committee (PAC) eventually will give to every political candidate, and in studies of criminal recidivism, the models do not assume that all former prisoners eventually will return to prison. Split-population models estimate the proportion of observations that will not experience the event, together with the parameters characterizing the hazard rate for the proportion experiencing the event. These models allow differential effects for the covariates on whether the event occurred and its timing. For example, a covariate may have a positive effect on whether a contribution is made but a negative effect on when it is made. The appeal of the log-likelihood in a split-population model is that observations that never experience the event contribute information only to part of the function. As such, the log-likelihood “splits” the two populations (Schmidt & Witte, 1988).

DIAGNOSTICS

As in the traditional regression setting, residual analysis in the event history analysis context is a method of checking specification or model adequacy. Various pseudo-residuals are defined in event history analysis for checking different aspects of a model, taking into account the complication that censoring adds to the definition of a residual. In addition to their use in testing the proportional hazards assumption, pseudo-residuals can help the researcher in assessing the model fit, functional form of the covariates, and influence of particular observations. For example, the martingale residuals can be calculated to test whether a given covariate X should be entered linearly, as a quadratic, or in one of the many other possibilities. These diagnostic methods should be used routinely in applications to ensure the integrity of the model (Therneau & Grambsch, 2000).

SINGLE AND MULTIPLE EVENTS

In addition to studying a single-event occurrence, where once the event is experienced, the observation

leaves the risk set, event history analysis also can consider multiple events. Event history models for multiple events take into account the lack of independence across events, because ignoring the correlation can yield misleading variance estimates and possibly biased estimates of the coefficients.

Multiple events can be simultaneous unordered events whose risk of occurring varies. In this case, we consider one of several types of “failure,” and such processes are referred to as *competing risks*. For example, a member of the U.S. House of Representatives may leave the House in a variety of substantively interesting ways that we should recognize and incorporate into our models, such as being defeated in the primary, being defeated in the general election, running for higher office, or retiring. We would expect that the hazard rate and effects of the covariates will differ across these types of departure.

One can also consider ordered multiple (or *repeated*) events. Repeated-events processes, in which subjects experience the same type of events more than once, are common in fields as diverse as public health, criminology, labor and industrial economics, demography, and political science. Failing to account for repeated events implicitly assumes that the first, second, third, and subsequent events are statistically independent of one another, a strong and usually untenable assumption. The conditional interevent (or gap) time model will be applicable for most instances of repeated events in social science. However, the nature of the means by which repeated events occur (that is, sequentially or simultaneously) and the corresponding construction of the “risk set” for each observation should provide the primary motivation for selecting one model over another (Box-Steffensmeier & Zorn, 2002).

EXAMPLE AND INTERPRETATION

Many applications of event history analysis exist in the social sciences, and the substantive realm of problems being studied is greatly expanding. Examples include studies of the duration of unemployment, peace, survey response time, criminal recidivism, marriages, public policy program implementation, and lobbying. Table 1 presents typical Cox proportional hazard estimates, using militarized conflict data. (Efron’s approximation for ties is used in the estimation of the Cox model. The Breslow approximation was the first approximation developed and is not generally

Table 1 Cox Proportional Hazards Model of Conflict

	β (S.E.)	p value
Democracy	-0.439(0.100)	< 0.001
Growth	-3.227(1.229)	0.009
Alliance	-0.414(0.111)	< 0.001
Contiguous	1.213(0.121)	< 0.001
Capability ratio	-0.214(0.051)	< 0.001
Trade	-13.162(10.327)	0.202
Wald or LR test	272.35 ($df = 6$)	< 0.001

NOTE: $N = 20,448$.

recommended, whereas the exact likelihood option typically gives results extremely similar to those obtained with the Efron approximation while taking considerably longer to converge.)

Oneal and Russett's (1997) widely used data on the relationship among economic interdependence, democracy, and peace is used for the illustration. The data consist of 20,448 observations on 827 dyads (i.e., pairs of states such as the United States and Canada), between 1950 and 1985. We model the hazard of a militarized international conflict as a function of six primary covariates (some of which vary over time): a score for *democracy* (a dyadic score for the two countries which ranges from -1 to 1), the level of *economic growth* (the lesser rate of economic growth, as a percentage, of the two countries), the presence of an *alliance* in the dyad (a dummy variable indicating whether the two countries were allied), the two nations' *contiguity* (a dummy variable for geographic contiguity), their military *capability ratio* (a ratio measuring the dyadic balance of power), and the extent of bilateral *trade* in the dyad (a measure of the importance of dyadic trade to the less trade-oriented country; it is the ratio of dyadic trade to the gross domestic product of each country). (See Oneal and Russett, 1997, for details of the variables and coding.)

Liberal theory suggests that all variables except contiguity ought to decrease the hazard of a dispute, while contiguity should increase it. The likelihood ratio (LR) test at the bottom of Table 1 shows that the specified model is preferred to the null model (i.e., the null hypothesis is that there is no statistically significant difference between the specified model and the null model of no independent variables). All the coefficients are in the expected directions, and all except that for trade are statistically significant. Note that the Cox model does not have an intercept term; it is absorbed

into the baseline hazard. Because the coefficients of the Cox model are parameterized in terms of the hazard rate, a positive coefficient indicates that the hazard is increasing, or "rising," with changes in the covariate (and hence survival time is decreasing), and a negative sign indicates the hazard is decreasing as a function of the covariate. For this model, the negative coefficient of -0.439 for democracy suggests that dyadic democracy reduces the likelihood of conflict; that is, dyadic democracy results in a lower hazard (and longer survival time). Box-Steffensmeier and Jones (1997) used the percentage change in the risk of experiencing the event to understand the impact of the effect (p. 1434). For a dichotomous independent variable, the percentage change in the risk of experiencing the event is

$$100[e^{(\beta_k \times 1)} - e^{(\beta_k \times 0)}]/e^{(\beta_k \times 0)}.$$

Negative coefficients produce values of $e^{(\beta_k \times 1)}$ that are less than one, and therefore produce negative percentage changes. The interpretation for a continuous independent variable is similar:

$$100[e^{[\beta_k \times (x+\delta)]} - e^{(\beta_k \times x)}]/e^{(\beta_k \times x)}.$$

This gives the percentage change in the hazard rate for a δ -unit change in the independent variable, x . So, a one-unit increase in the democracy variable corresponds to a $[(e^{(-0.439)} - 1) \times 100] = 36\%$ decrease in the hazard of conflict at any given time.

In actuality, the militarized conflict data are characterized by large numbers of repeated events; for example, Britain and Germany fought each other in both World War I and World War II. Box-Steffensmeier and Zorn (2002) used these data to illustrate repeated events duration modeling and show that important differences are uncovered by taking into account the dependence generated from repeated conflicts.

Social science theories are increasingly focused on change processes, and temporal data are becoming widely available. Event history analysis is ideally suited for leveraging these research elements. The flexibility of the techniques, recent extensions for multiple events, and the incorporation of the observation's history about the events of interest are all compelling reasons to expect the use and popularity of event history techniques to increase in the social sciences.

—Janet M. Box-Steffensmeier

REFERENCES

- Blossfeld, H.-P., & Rohwer, G. (2002). *Techniques of event history modeling: New approaches to causal analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Box-Steffensmeier, J. M., & Jones, B. S. (1997). Time is of the essence: Event history models in political science. *American Journal of Political Science*, 41, 336–383.
- Box-Steffensmeier, J. M., & Zorn, C. J. W. (2001). Duration models and proportional hazards in political science. *American Journal of Political Science*, 45, 951–967.
- Box-Steffensmeier, J. M., & Zorn, C. J. W. (2002). Duration models for repeated events. *The Journal of Politics*, 64(4), 1069–1094.
- Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34, 187–220.
- Hald, A. (1990). *A history of probability and statistics and their applications before 1750*. New York: John Wiley and Sons.
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.
- Oneal, J. R., & Russett, B. (1997). The classical liberals were right: Democracy, interdependence, and conflict, 1950–1985. *International Studies Quarterly*, 41, 267–294.
- Schmidt, P., & Witte, A. D. (1988). *Predicting recidivism using survival models*. New York: Springer-Verlag.
- Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. New York: Springer-Verlag.

EVENT SAMPLING

Event sampling refers to a diverse class of specific empirical methods for studying individual experiences and social processes within their natural, spontaneous context. Event sampling procedures are designed to obtain reasonably detailed accounts of thoughts, feelings, and behaviors as they occur in everyday life. Examples include the experience sampling method (ESM), in which respondents are signaled at random moments during the day and asked to describe their activity at that exact moment, and daily diaries, in which at the end of each day, for some period (typically ranging from 1 week to 1 month), respondents report their experiences on that day. In one typical ESM study, 50 college students were signaled by pagers seven times per day for 1 week. When signaled, students with higher need for intimacy (as assessed on a projective personality test) were more likely to be thinking about people and relationships, were more often engaged in conversations with others and less likely to wish to be alone, and reported more positive affect if they were socializing (McAdams & Constantian, 1983).

Event sampling methods require that participants monitor and describe their ongoing activity along dimensions and according to schedules and formats defined by the researcher. Event sampling has three fundamental rationales that differentiate it from other common research paradigms (e.g., laboratory experiments, surveys): (a) that because behavior is influenced by context, it is important to sample behavior in its natural environment; (b) that global, retrospective reports are often biased by people's limited abilities to remember and summarize numerous events over time; and (c) that accounts of seemingly ordinary, everyday experience, when properly examined, are capable of providing valuable insights about human behavior. Topics for which event sampling methods have been employed profitably include emotion, social interaction, pain, smoking, stress and coping, student motivation, exercise, eating disorders, psychopathology, self-relevant cognition, personality, intergroup relations, and evaluations of drug treatments and therapeutic interventions.

Most event sampling studies employ one of three general protocols: *time-contingent responding*, in which participants report their experiences at fixed intervals (e.g., daily, hourly); *event-contingent responding*, in which a report is solicited whenever a predefined event occurs (e.g., smoking a cigarette, conversing with a friend), and *signal-contingent responding*, in which participants describe their experiences when signaled to do so by some device (e.g., a pager or a preprogrammed portable computer). Signals may follow a fixed or random schedule. Although event sampling was originated with simple paper-and-pencil responses, recent developments in electronic recording devices [e.g., personal digital assistants (PDAs) and voice recorders], as well as in ambulatory physiological monitoring, have added considerably to the validity, flexibility, and range of these methods.

Event sampling is not limited to participant self-reports. For example, in an observational study of schoolchildren, observers might code the behavior of a target child (e.g., what the child is doing, with whom he or she is currently interacting, visible affective expressions, and so on) according to any of the above schedules (e.g., every 10 minutes, after conflict, or following a randomized schedule, respectively).

In a typical event sampling study, a researcher might obtain a series of detailed descriptions of adolescents' momentary moods and actions across a 2-week period. These records could then be used in several ways: for

example, to determine how mood differs depending on what the adolescent is doing and with whom he or she is doing it, to isolate personality factors or other individual differences associated with differential patterns of mood or activity, or to predict outcomes, such as school grades, social competence, or health, from a summary of these momentary reports. Regardless of the particulars, event sampling researchers share a fascination with observing ordinary behavior in its natural context and a conviction that such observations hold important clues about key questions in the behavioral and social sciences.

—Harry T. Reis

REFERENCES

- Bolger, N., Davis, A., & Rafaeli, E. (in press). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*.
- Hektner, J. M., & Csikszentmihalyi, M. (2002). The experience sampling method: Measuring the context and the content of lives. In R. B. Bechtel & A. Churchman (Eds.), *Handbook of environmental psychology* (pp. 233–243). New York: John Wiley & Sons.
- McAdams, D. P., & Constantian, C. A. (1983). Intimacy and affiliation motives in daily living: An experience sampling analysis. *Journal of Personality and Social Psychology*, 45, 851–861.
- Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 190–222). New York: Cambridge University Press.
- Stone, A. A., Shiffman, S., & DeVries, M. (1999). Ecological momentary assessment. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 27–38). New York: Russell Sage.

EXOGENOUS VARIABLE

An exogenous variable is a factor in CAUSAL MODELING or causal system whose value is independent from the states of other variables in the system; that is, it is a factor whose value is determined by factors or variables outside the causal system under study. For example, rainfall is exogenous to the causal system constituting the process of farming and crop output. There are causal factors that determine the level of rainfall—so rainfall is endogenous to a weather model—but these factors are not themselves part of the causal model one would use to explain the level of crop output.

As with endogenous variables, the status of the variable is relative to the SPECIFICATION of a particular model and causal relations among the INDEPENDENT VARIABLES. An exogenous variable is by definition one whose value is wholly causally independent from other variables in the system. The category of “exogenous” variables therefore is contrasted to those of “purely endogenous” and “partially endogenous” variables. A variable can be made endogenous by incorporating additional factors and causal relations into the model.

There are causal and statistical interpretations of exogeneity. The causal interpretation is primary and defines exogeneity in terms of the factor’s causal independence from the other variables included in the model. The statistical or ECONOMETRIC concept emphasizes noncorrelation between the exogenous variable(s) and the error term(s) in the model. Typical REGRESSION models assume that all the independent variables are exogenous.

—Daniel Little

See also ENDOGENOUS VARIABLE

REFERENCES

- Engle, R. F., Hendry, D. F., & Richard, J. F. (1983). Exogeneity. *Econometrica*, 51, 277–304.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.
- Woodward, J. (1995). Causation and explanation in econometrics. In D. Little (Ed.), *On the reliability of economic models: Essays in the philosophy of economics* (pp. 9–61). Boston: Kluwer Academic.

EXPECTANCY EFFECT. See EXPERIMENTER EXPECTANCY EFFECT

EXPECTED FREQUENCY

An expected *value* equals the average or mean of some statistic, whereas an expected *frequency* is a mean count or mean number of times an event occurs. Events could be a category of a single discrete variable, a cell of a cross-classification of two or more discrete

variables, or some other specified event (e.g., the number of goals scored by a team during a soccer game). An expected frequency is computed by multiplying the probability that an event occurs by the total number of possible times that the event could occur. For example, consider random samples of size $n = 75$ people from a population in which the probability that an individual is left-handed equals $\pi = 0.10$. The number of left-handers in a sample could be any integer between 0 and 75; however, the expected or mean number of left-handers equals $n\pi = 75(0.10) = 7.5$. Although frequencies are always integers, expected frequencies typically are not integers.

Expected frequencies are used in test statistics for assessing hypotheses and model goodness of fit. Two common test statistics are Pearson's CHI-SQUARE TEST, $\chi^2 = \sum_{i=1}^N (f_i - F_i)^2 / F_i$, and the LIKELIHOOD RATIO STATISTIC, $L^2 = \sum_{i=1}^N f_i \log(f_i / F_i)$, where f_i equals the observed number of observations for category i , F_i equals the expected frequency for category i , and N equals the total number of categories. If a null hypothesis or model for data specifies a value for the probability (or probabilities), such as "the probability of being left-handed is 0.10," then an expected frequency is computed as $F_i = n\pi_i$, where n equals the total number of observations.

The probabilities of events usually are unknown and are estimated from data by computing sample proportions or by fitting a model to data. When data are used to estimate probabilities, the resulting expected frequencies are known as estimated expected frequencies. For example, consider the data in Table 1, which is reformatted from data presented by Jon Cohen (2001). The data consist of the number of smallpox cases in Liverpool, England, during 1902–1903, cross-classified according to whether a person was vaccinated in infancy and whether the person died of smallpox. Assuming that vaccinations do not affect the severity of a case of smallpox, the probability of death as a result of smallpox is estimated by computing the proportion of smallpox cases that resulted in death, that is, the total number of deaths divided by the total number of cases, which equals $\hat{\pi}_{\text{death}} = 79/1,163 = 0.0679$. The estimated expected frequency of death given vaccination equals $\hat{F}_{\text{vac,death}} = n_{\text{vac}}\hat{\pi}_{\text{death}} = 943(0.0679) = 64.0559$. The estimated expected frequencies for other cells in Table 1 could also be computed. For a hypothesis test, the estimated expected frequencies, $\hat{F}_i$, replace the expected

Table 1 Smallpox Cases in Liverpool, England (1902–1903)

	<i>Death Due to Smallpox</i>		<i>Total</i>
	<i>Yes</i>	<i>No</i>	
Vaccinated in infancy	19	924	943
Not vaccinated	60	160	220
Total	79	1,084	1,163

SOURCE: Cohen (2001, p. 985).

frequencies, F_i , in Pearson's chi-square statistic or the likelihood ratio statistic.

—Carolyn J. Anderson

REFERENCES

- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.
- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.
- Cohen, J. (2001). Smallpox vaccinations: How much protection remains? *Science*, 294, p. 985.

EXPECTED VALUE

An expected value is the long-run average of a RANDOM VARIABLE X . More formally, the expected value $E(X)$ is a weighted average of all of X 's possible values. For a DISCRETE variable, the weights are the PROBABILITIES of the X values, $p(X)$, and the average is obtained by summation:

$$E(X) = \sum_{x=-\infty}^{\infty} xp(x).$$

For a CONTINUOUS VARIABLE, the weights are the probability densities of the X values, $f(X)$, and the average is obtained by integration:

$$E(X) = \int_{x=-\infty}^{\infty} xf(x).$$

For certain distributions, the integral or sum fails to converge. In such cases, the expectation is undefined. Although undefined expectations are fairly rare, a well-known example occurs in connection with the

Cauchy distribution, which is defined as the ratio of two standard normal variables.

To illustrate the calculations, consider the flipping of a fair coin. Let X be 1 if the coin comes up heads, and 0 if it comes up tails. Each value of X has the same probability, $p(X = 1) = p(X = 0) = 0.5$, so the expected value, using the discrete variable formula, is

$$E(X) = 1(0.5) + 0(0.5) = 0.5.$$

Whenever X is a DUMMY VARIABLE, as here, the expected value is the probability that X is 1. When X is an INTERVAL variable, the expected value is the population mean.

Expectation is a linear operation: The expected value for a weighted sum of two random variables is just the weighted sum of the expectations,

$$E(aX + bY) = aE(X) + bE(Y),$$

where a and b are the weights and X and Y are the variables. For example, suppose that X and Y are dummy variables associated with two fair coins, and each variable is 1 if its coin comes up heads. If both coins are tossed, the expected number of heads is

$$E(X) + E(Y) = 0.5 + 0.5 = 1.$$

Just as the expectation of a sum is the sum of the expectations, the expectation of a product is just the product of the expectations:

$$E(XY) = E(X)E(Y).$$

The expectation for a general function g of X is just a weighted average of the values of $g(X)$. For a discrete variable X , the weights are the probabilities of the X values, $p(X)$, and the average is obtained by summation:

$$E(g(X)) = \sum_{x=-\infty}^{\infty} g(x)p(x).$$

For a continuous variable X , the weights are the densities of the X values, $f(X)$, and the average is obtained by integration:

$$E(g(X)) = \int_{x=-\infty}^{\infty} g(x)f(x).$$

These formulas extend in a straightforward way to the multivariate case where X and Y are vectors of random variables, and g is a function, possibly vector-valued, of the variable X .

—Paul T. von Hippel

REFERENCE

Rice, J. A. (1995). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury Press.

EXPERIMENT

In social sciences, an experiment is a research strategy used by a social scientist to establish causal relationships between one or more independent variables and one or more dependent variables. An independent variable is a variable that is manipulated by a researcher; its causal impact on the dependent variable is investigated by the researcher. An independent variable is alternatively called the treatment variable or factor. The dependent variable is the observed phenomenon or measurement that has been affected by the manipulation of the independent variable. Dependent variables are alternatively referred to as outcome or criterion variables.

In establishing the causal relationship between independent variables and dependent variables, the researcher should design an experiment that includes the following elements:

- Manipulation of the amount (as in the case of quantitative independent variables) or the level of the independent variable (as in the case of qualitative independent variables)
- Control of nuisance (or confounding) variables using random selection and random assignment of subjects into treatment conditions
- Careful recording or observation of the change in the dependent variable

The first and the second requirements are achieved in either laboratory studies or field experiments (see FIELD EXPERIMENT); they distinguish the experimental research strategy from other research strategies such as quasi-experimental studies, surveys, or naturalistic studies. For these reasons, experiments that possess the above three characteristics are sometimes called true experiments (Campbell & Stanley, 1966).

An example of a true experiment is reported in Cordova and Lepper's (1996) study of elementary school children's learning of arithmetical order-of-operations rules. All learning activities took place in a computerized environment. First, children were randomly assigned, within gender, to either a control condition or one of four experimental conditions. In the

control condition, learning materials were presented abstractly. In the four experimental conditions, identical materials were contextualized in a meaningful and visually enhanced format. For half of the students in the four experimental conditions, the learning context was individually personalized; the other half were presented with a generic format. Furthermore, in each experimental condition, half of the group was given choices regarding instructional aspects of the learning context while the other half was not. Thus, the four experimental conditions were (a) learning context personalized with choices, (b) learning context personalized without choices, (c) learning context not personalized but with choices, and (d) learning context not personalized and also without choices. In sum, three independent variables were manipulated in the study: contextualization, personalization, and choice. The dependent variables were several, including students' intrinsic motivation for learning, depth of engagement during learning, the amount of learning acquired in a fixed period of time, students' perceived self-competence, and levels of aspiration. The causal relationship between three independent variables and multiple dependent variables was established through (a) random assignment of subjects into control and experimental conditions; (b) control of miscellaneous variables such as grade level of subjects, gender, and learning materials; and (c) statistical theory of inference making.

Indeed, inferential statistical theories enable social and behavioral scientists to make sense of empirical data collected in the so-called true experiments, in which causal relationships are established not by ruling out all possible alternative explanations or controlling all possible confounding variables, but by probabilities. The logic behind inferential statistical theories is testing statistical hypotheses, in the light of data, followed by making inferences about the underlying populations in the tradition of inductive logical reasoning (Maxwell & Delaney, 1990). The principles and statistical theories that have guided experimental research studies in social and behavioral sciences were, to a great extent, formulated by Karl Pearson (1857–1936), Sir Ronald A. Fisher (1890–1962), and Jerzy Neyman (1894–1981).

Three basic research designs have been employed routinely by social scientists to carry out true experiments or, simply, experiments (Kirk, 1995). They are the completely randomized design, RANDOMIZED-BLOCKS DESIGN, and LATIN SQUARE design. Kirk refers

to these as building blocks to all research designs used in experiments. Each is described below.

COMPLETELY RANDOMIZED DESIGN

This research design includes one independent variable and one dependent variable. The independent variable may have two or more levels (or treatment conditions), and the dependent variable is always continuous. If the Cordova and Lepper (1996) study cited earlier is simplified to include only one independent variable, say *contextualization*, and the dependent variable observed is *intrinsic motivation*, the simplified study design will be a bona fide completely randomized design provided that subjects are randomly assigned to either the contextualized or the abstract condition at the onset. A layout of this simplified study may look like Table 1:

Table 1

Subject	Independent Variable	
	Contextualized Condition	Abstract Condition
S_1	Y_{11}	Y_{12}
S_2	Y_{21}	Y_{22}
.	.	.
.	.	.
.	.	.
S_{15}	$Y_{1,15}$	$Y_{2,15}$

The statistical linear model for the data collected in a completely randomized design is

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \quad (i = 1, \dots, n \text{ or } 15 \text{ in this case; } j = 1, \dots, p \text{ or } 2 \text{ in this case}),$$

where

- Y_{ij} is the i th subject's score in the j th treatment condition of the independent variable
- μ is the grand mean, therefore a constant, of the population of all subjects
- α_j is the treatment effect of the j th treatment condition of the independent variable; algebraically, it equals the deviation of the population mean (μ_j) from the grand mean (μ). It is a constant for all subjects' dependent scores in the j th condition, subject to the constraint that all α_j sum to zero across all treatment conditions of the independent variable

ε_{ij} is the random error effect associated with the i th subject randomly assigned to the j th level of the independent variable; it is a random variable that is normally distributed in the underlying population and is independent of the effect α_j

The null (H_0) and alternative (H_1) hypotheses for the completely randomized design are

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_p = 0,$$

$$H_1: \text{at least one } \alpha_p \neq 0.$$

To test the null hypothesis, data are organized to form a ratio of MEAN SQUARES between (MS_b) over mean squares within (MS_w). The ratio is distributed as an F distribution provided that the normality, homogeneity of variance, and independence assumptions about the data are met. The normality assumption states that the underlying distribution of data in all treatment conditions of the independent variable is the normal distribution. The homogeneity of variance assumption is that all normal population distributions have identical variances. Finally, the independence assumption says that random error effects, ε_{ij} , are independent from each other both within and across treatment conditions. Failure to meet these assumptions results in quasi F ratios that may inflate the Type I error rate (Glass, Peckham, & Sanders, 1972). The example given above is obviously the simplest completely randomized design, with only two treatment conditions of the independent variable. In this case, the test statistic can be either an F or t , as in the comparison of two independent group means. Most applications of the completely randomized design in social sciences involve three or more treatment conditions.

If treatments of the independent variable included in a study (e.g., contextualized and abstract) are the only treatments that a researcher is interested in investigating and, consequently, making generalizations about, the corresponding independent variable is regarded as a fixed factor, and the design is a FIXED-EFFECTS DESIGN. The statistical model prescribed for a fixed-effects design is MODEL I (Kirk, 1995). If, however, treatments included in a study represent only a sample of all treatments that a researcher wishes to study, the corresponding independent variable is treated as a random factor. The design is a random-effects design, and

MODEL II (ANOVA) is an appropriate statistical model for the data (Kirk, 1995).

RANDOMIZED-BLOCKS DESIGN

The use of completely randomized design requires that subjects be randomly assigned into treatments of an independent variable. Hence, subjects in a treatment condition are considered an independent sample (or subsample) from subjects in another treatment condition. This characteristic of independence is factored into the calculation of the F or t statistic when the statistical test is performed. For many studies, such as those involving repeated measurements of the same subjects, it is not possible to maintain the independence among subjects in different treatment conditions. For other research, it is desirable (as in family or marital research) to pair up subjects (such as husbands and wives or classmates) so that researchers can study the dynamics of the relationship between subjects. Researchers may purposely introduce dependence by matching subjects on certain characteristics that are related to the dependent measures to ensure that these characteristics are “controlled” in the study. For example, in the simplified Cordova and Lepper (1996) study described in the previous section, one might wish to match subjects on prior knowledge that is related to the dependent measure (i.e., *intrinsic motivation*). By so doing, the source of variance that is due to prior knowledge is planned into the study and, therefore, can be isolated from the source of variance that is due to the treatment (i.e., *contextualization*). Let’s further assume that prior knowledge is differentiated at three levels: high, medium, and low. An appropriate research design to control for prior knowledge is the randomized-blocks design. Table 2 illustrates the design features.

The variable Prior Knowledge is technically called a blocking variable; its levels (high, medium, and low) are blocks. Within each block of the blocking variable, subjects (e.g., $S_1, \dots, S_5$) are homogeneous insofar as prior knowledge is concerned. Subjects in each block are randomly assigned to treatment conditions (e.g., contextualization condition or abstract condition). The random assignment of subjects to treatments within each block and the planned differences across blocks ensure that the idiosyncratic differences of subjects are distributed randomly among treatment conditions, and differences on the blocking variable are equally

Table 2

Prior Knowledge	Independent Variable		
	Subject	Contextualized Condition	Abstract Condition
High	S_1	Y_{11}	Y_{12}
	.	.	.
	.	.	.
	.	.	.
	S_5	Y_{15}	Y_{25}
Medium	S_6	Y_{16}	Y_{26}
	.	.	.
	.	.	.
	.	.	.
	S_{10}	$Y_{1,10}$	$Y_{2,10}$
Low	S_{11}	$Y_{1,11}$	$Y_{2,11}$
	.	.	.
	.	.	.
	.	.	.
	S_{15}	$Y_{1,15}$	$Y_{2,15}$

represented in each treatment condition. Thus, results are not systematically biased as a result of differences in the blocking variable. In this sense, the randomized-blocks design can isolate and attribute a portion of score variance to the blocking variable, whereas the completely randomized design cannot.

The statistical model assumed for a randomized-blocks design is

$$Y_{ij} = \mu + \alpha_j + \pi_i + \varepsilon_{ij} \quad (i = 1, \dots, n \text{ or } 3 \text{ in the example above; and } j = 1, \dots, p \text{ or } 2 \text{ in the example}),$$

where

- Y_{ij} is the score of the i th block and the j th treatment condition
- μ is the grand mean, therefore a constant, of the population of observations
- α_j is the treatment effect for the j th treatment condition; algebraically, it equals the deviation of the population mean (μ_j) from the grand mean (μ or $\mu_{..}$); it is a constant for all observations' dependent score in the j th condition, subject to the restriction that all α_j sum to zero across all treatment conditions as in a fixed-effects design, or the sum of all α_j equals zero in the populations of treatments, as in the random-effects model

π_i is the block effect for population i and is equal to the deviation of the i th population mean ($\mu_{i.}$) from the grand mean ($\mu_{..}$); it is a constant for all observations' dependent score in the i th block, normally distributed in the underlying population

ε_{ij} is the random error effect associated with the observation in the i th block and j th treatment condition; it is a random variable that is normally distributed in the underlying population of i th block and j th treatment condition and is independent of π_i .

The null (H_0) and alternative (H_1) hypotheses for the randomized-blocks design are identical to those for the completely randomized design. To test the null hypothesis, data are organized to form a ratio of mean squares treatment (MS_t) over mean squares residual (MS_r). Mean squares residual is a ratio of the sum of squares residual over its degrees of freedom, where the sum of squares residual is the difference of sum of squares within (SS_{within}) minus the sum of squares due to the blocking variable. The ratio is distributed as an F distribution provided that the normality, homogeneity of variance, and independence assumptions about the data are met.

The example given above is the simplest randomized block design, with only two treatment conditions of the independent variable. In this case, the test statistic can be either an F or t , as in the comparison of two dependent (or matched pair) group means. Most applications of the randomized block design in the social sciences involve three or more treatment conditions.

If treatments of the independent variable included in a study (e.g., contextualized and abstract) are regarded as fixed and the blocking variable is regarded likewise, the design is a fixed-effects design. The statistical model prescribed for a fixed-effects design is Model I (Kirk, 1995). If, however, the blocking variable is regarded as a RANDOM FACTOR, then the corresponding design is a MIXED DESIGN and the model is MODEL III (ANOVA). If treatments included in a study represent only a sample of all treatments, and hence the corresponding independent variable is a random factor, the design is a random-effects design and Model II is an appropriate statistical model for the data (Kirk, 1995).

When more than one subject is assigned to the combination of block and treatment condition, the design is referred to as the generalized randomized block design. This is the case with the example given here.

Alternatively, the same subject (or observation) can serve as a block and is subject to all treatments in a random sequence. Some authors prefer to call this specific variation of block designs a REPEATED-MEASURES DESIGN. Details on how the blocks are formed and on additional statistical assumptions uniquely associated with this design and its corresponding statistical test are found in the entries titled “Block Designs” and “Randomized-Blocks Design.”

LATIN SQUARE DESIGN

The Latin square design is an experimental design or QUASI-EXPERIMENT that can isolate and attribute a portion of score variance to two blocking variables, whereas the randomized block design controls for only one blocking variable.

Much like the Rubik’s cube, all Latin square designs are square (2 by 2, 3 by 3, 4 by 4, etc.); that is, the two blocking variables and the independent variable all have the same number of levels (or conditions). Let us assume that the second blocking variable we wish to control in the simplified Cordova and Lepper study described previously is gender, which has two levels. The independent variable also has two levels: contextualized and abstract. Consequently, the first blocking variable, prior knowledge, has to be redefined to comprise only two levels: high and low. Table 3 illustrates this 2 by 2 Latin square design.

Table 3

Prior Knowledge	Gender	
	Female	Male
High	Contextualized condition	Abstract condition
Low	Abstract condition	Contextualized condition

Notice that in Table 3, the treatments are shown inside the square, while the two blocking variables are presented as row and column variables. Indeed, treatment conditions occur only once in each row and once in each column.

The statistical model for the above Latin square design is

$$Y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_l + \varepsilon_{\text{pooled}},$$

($i = 1, \dots, n$ in the example above; $j, k,$
and $l = 1, \dots, p$ or 2 in the example),

where

Y_{ijkl}	is the score of the i th subject in the j th treatment condition, k th level of the first block variable (e.g., prior knowledge) and the l th level of the second blocking variable (e.g., gender)
μ	is the grand mean, therefore a constant, of the population of observations
α_j	is the treatment effect for the j th treatment condition; algebraically, it equals the deviation of the population mean ($\mu_{j..}$) from the grand mean (μ or $\mu_{...}$); it is a constant for all subjects’ dependent score in the j th condition, subject to the restriction that all α_j sum to zero across all treatment conditions as in a fixed-effects design, or the sum of all α_j equals zero in the populations of treatments, as in the random-effects model
β_k	is the block effect for the k th population and is equal to the deviation of the k th population mean ($\mu_{.k.}$) from the grand mean ($\mu_{...}$); it is a constant for all subjects’ dependent score in the k th block, normally distributed in its underlying population
γ_l	is the block effect for the l th population and is equal to the deviation of the l th population mean ($\mu_{..l}$) from the grand mean ($\mu_{...}$); it is a constant for all subjects’ dependent score in the l th block, normally distributed in its underlying population
$\varepsilon_{\text{pooled}}$	is the pooled error effect associated with Y_{ijkl} ; algebraically, this concept equals the deviation of the observed score Y_{ijkl} from all other parameters in the model (i.e., $\mu, \alpha_j, \beta_k,$ and γ_l) and, by definition, it is statistically independent of all the parameters in the model

The null (H_0) and alternative (H_1) hypotheses for the Latin square design continue to be identical to those for the completely randomized design. To test the null hypothesis, data are organized to form a ratio of mean squares treatment (MS_t) over mean squares pooled error (MS_{pooled}). Mean squares pooled error is a ratio, consisting of the sum of squares pooled error over its degrees of freedom, where the sum of squares pooled error is the difference of the sum of squares within (SS_{within}) minus the sum of squares due to *two* blocking variables. The ratio is distributed as an F distribution provided that the normality, homogeneity of variance, and independence assumptions about the

data are met. Furthermore, the F test assumes that no interaction exists among the two blocking variables and the independent variable (also the treatment). To test this interaction assumption, a Latin square design must have two or more observations per cell (i.e., $n \geq 2$). Otherwise, this assumption of no interaction is assumed for data.

If treatments of the independent variable included in a study (e.g., contextualized and abstract) are regarded as fixed and the two blocking variables are regarded likewise, the design is a fixed-effects design. The statistical model prescribed for a fixed-effects design is Model I (Kirk, 1995). If, however, one of the blocking variables is regarded as a random factor, the corresponding design is a mixed-effects design and the model is Model III. If treatments included in a study represent only a sample of all treatments, and hence the corresponding independent variable is a random factor, the design is a random-effects design and Model II is an appropriate statistical model for the data. For additional readings on Latin square designs, readers are encouraged to consult the entry titled "Latin Square Design," Kirk (1995), or Maxwell and Delaney (1990).

For the three research designs and the corresponding statistical models presented above, it is assumed that data are collected and measured without errors. If measurement errors are present in the data, RELIABILITY needs to be taken into consideration when the F statistic is computed (Cleary & Linn, 1969; Levin & Subkoviak, 1977). Furthermore, the VALIDITY of any causal relationship established between independent variables and one or more dependent variables in an experiment can be threatened by a number of factors. For details on threats to external, internal, and statistical conclusion validities, readers are encouraged to consult Campbell and Stanley (1966) and Kirk (1995).

Within the field of probability theory, an EXPERIMENT is a process or procedure that leads to a single outcome whose probability of occurrence is to be determined. An experiment can be a laboratory study (e.g., engaging students in collaborative learning to improve their social skills) or a process of observing an aspect of behavior in a sample taken from a population (e.g., noting the birth months of two randomly chosen guests at a dinner party of 50). Both the experiment and the outcome (the improvement of social skills or two guests born in the same month) should be well defined such that a researcher could formulate rules for determining the probability (likelihood) of obtaining the specific outcome, among all possible outcomes. The

approach to probability determination may be based on either (a) the subjective-personalistic framework, (b) the logical or classical framework, or (c) the empirical-relative or frequency framework. In this context, an experiment may also be called a simple experiment.

—Chao-Ying Joanne Peng

See also BLOCK DESIGN, FIELD EXPERIMENTATION, LATIN SQUARE, QUASI-EXPERIMENT, RELIABILITY, VALIDITY

REFERENCES

- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally College Publishing.
- Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *The British Journal of Mathematical and Statistical Psychology*, 22(1), 49–55.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally College Publishing.
- Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88(4), 715–730.
- Fisher, R. A. (1971). *Design of experiments*. New York: Hafner Press. (Original work published 1935.)
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the analysis of variance and covariance. *Review of Educational Research*, 42, 237–288.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.
- Levin, J. R., & Subkoviak, M. J. (1977). Planning an experiment in the company of measurement error. *Applied Psychological Measurement*, 1(3), 331–338.
- Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.
- Vogt, W. P. (1999). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (2nd ed.). Thousand Oaks, CA: Sage.

EXPERIMENTAL DESIGN

CHARACTERISTIC FEATURES OF EXPERIMENTS

Two characteristics set experimentation apart from other methods of social inquiry. First, experimentation involves a planned intervention. Unlike passive

observation, in which the researcher attempts to trace the effects of naturally occurring fluctuations in putative causes, experiments create a disturbance in the environment and track its consequences. For example, an observational study of welfare reform might examine whether rates of joblessness vary with over-time or cross-jurisdiction variations in program requirements. By contrast, an experimental investigation would assign program participants to different sets of requirements in order to examine whether they affected joblessness.

The second characteristic of experiments is RANDOM ASSIGNMENT. Although the term EXPERIMENT is used loosely in common parlance to refer any type of intervention, in social science it has come to refer to studies in which units of observation are assigned at random to treatment and control conditions. Because social scientists cannot create the equivalent of a physics lab, in which all extraneous causes are eliminated, they instead rely on random assignment, which, as R. A. Fisher pointed out in his classic work *The Design of Experiments* (1935), is the only *procedure* that guarantees the comparability of treatment and control groups. So long as randomization is carried out faithfully, we can be sure that treatment and control groups differ solely as a result of chance.

Randomization therefore enables researchers to make precise statistical statements about the likelihood that any postintervention differences between treatment and control groups resulted from fortuitous differences between the groups, as opposed to the intervention in question. Precise probability statements typically are unavailable in non-experimental social science, where the independent variables are generated by an unknown process. Beyond the usual statistical uncertainty that arises from limited sample size, non-experimental research involves considerable uncertainty about whether the observed correlation between two variables, X and Y , reflects the causal influence of X on Y , of Y on X , or of some third variable Z on both X and Y .

PROBLEMS CONFRONTING EXPERIMENTAL DESIGN

Although, in principle, experiments represent the most reliable means of drawing causal inferences, in practice experiments confront several problems. Problems of INTERNAL VALIDITY arise when an experiment fails to isolate the true effects of a particular

causative agent. For example, the Lanarkshire milk experiment, an early study designed to test whether distributing milk in schools improved students' growth rates, was undone when teachers reassigned underweight students in the control group to receive milk supplements. Thus, the contrast between treatment and control groups reflected both the effects of milk and the vagaries of teacher reassignments.

Concerns about EXTERNAL VALIDITY arise when the environment within which an experiment takes place or the people who participate in the experiment differ in important respects from the places and populations about whom the researcher intends to generalize. This concern arises frequently when LABORATORY EXPERIMENTS attempt to simulate economic or political environments using college students as subjects. In response to concerns about external validity, social scientists have turned increasingly to FIELD EXPERIMENTS, or experiments conducted in real-world settings. Such studies have examined the effectiveness of voter mobilization campaigns, preschool education programs, methods for encouraging compliance with tax rules, and a wide array of other interventions. Nevertheless, practical constraints limit the range of field experiments. Social scientists lack the resources and authority to manipulate large-scale causative factors, such as legislative institutions or the religiosity of the population.

Related to concerns about external validity is the issue of homogeneous treatment effects. Are all subjects equally influenced by a given intervention? If the answer is yes, the range of research opportunities expands, because one may study situations in which only some of those assigned to the treatment group actually receive treatment. For example, if a voter mobilization campaign has the same effect on everyone it contacts, but it reaches only half of the people it seeks to contact, its effect is twice as strong as a naive comparison of treatment and control groups would suggest. The assumption of homogeneous treatment effects, in other words, enables us to extrapolate easily from those who were actually exposed to a treatment to those whom researchers sought to treat.

Whether treatment effects are in fact homogeneous is an empirical question, not unlike the question of whether college students are as susceptible to social influences as those outside the university. Concerns about homogeneity and external validity underscore the importance of REPLICATING experimental findings in different settings and populations.

Ethical limitations are a further constraint. Experimental interventions may adversely affect the subjects involved, as well as the broader society. Besides the practical constraints of devising feasible interventions, researchers much follow procedural safeguards to ensure that subjects are neither harmed nor coerced. These considerations generally remove from consideration far-reaching experiments involving social or economic policy, but as Donald T. Campbell (1969) pointed out, the social costs of *not* conducting experiments must also be taken into account. Governments, firms, and organizations continually intervene in the world, and the question is whether their interventions can be structured in such a way as to generate useful knowledge.

EXPERIMENTS AND SCIENTIFIC ADVANCEMENT

Despite these limitations, experiments remain the gold standard for adjudicating causal claims. A single well-crafted experiment—conducted in a real-world setting and of sufficient size to produce statistically precise conclusions—can overshadow a large body of research based on observational data.

Part of the allure of experiments is their elegant transparency. In contrast to non-experimental data analysis, the analysis of experimental results often requires little more than elementary statistical methods, and the choice of statistical techniques seldom has a material bearing on the results. The experimental design largely dictates the manner in which the data will be analyzed statistically; committing to a plan of analysis *ex ante* helps guard against post hoc decisions that may bias the results in a particular direction. Experimental procedures not only lead to clearer causal inferences but also free the analyst from the moral hazards of data mining.

—Donald P. Green

See also EXPERIMENT

REFERENCES

- Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, 24, 409–429.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.
- Fisher, R. A. (1935). *The design of experiments*. New York: Hafner Publishing.

- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wilmington, MA: Houghton Mifflin.

EXPERIMENTER EXPECTANCY EFFECT

The experimenter expectancy effect is one of the sources of artifact or error in scientific inquiry (see INVESTIGATOR EFFECT). Specifically, it refers to the unintended effect of experimenters' hypotheses or expectations on the results of their research.

Some expectation of how the research will turn out is virtually a constant in science. Social scientists, like other scientists generally, conduct research specifically to examine hypotheses or expectations about the nature of things. In the social and behavioral sciences, the HYPOTHESIS held by the investigator can lead him or her unintentionally to alter behavior toward the research participants in such a way as to increase the likelihood that participants will respond so as to confirm the investigator's hypothesis or expectations. We are speaking, then, of the investigator's hypothesis as a self-fulfilling prophecy: One prophesies an event, and the expectation of the event then changes the behavior of the prophet in such a way as to make the prophesied event more likely. The history of science documents the occurrence of this phenomenon with the case of the horse Clever Hans as a prime example (Pfungst, 1911/1965).

The first experiments designed specifically to investigate the effects of experimenters' expectations on the results of their research employed human research participants. Graduate students and advanced undergraduates in the field of psychology were employed to collect data from introductory psychology students. The experimenters showed a series of photographs of faces to research participants and asked participants to rate the degree of success or failure reflected in the photographs. Half the experimenters, chosen at random, were led to expect that their research participants would rate the photos as being of more successful people. The remaining half of the experimenters were given the opposite expectation—that their research participants would rate the photos as being of less successful people. Despite the fact that all experimenters were instructed to conduct a perfectly standard

experiment, reading only the same printed instructions to all their participants, those experimenters who had been led to expect ratings of faces as being of more successful people obtained such ratings from their randomly assigned participants. Those experimenters who had been led to expect results in the opposite direction tended to obtain results in the opposite direction.

These results were replicated dozens of times employing other human research participants. They also were replicated employing animal research subjects. In the first of these experiments, experimenters were employed who were told that their laboratory was collaborating with another laboratory that had been developing genetic strains of maze-bright and maze-dull rats. The task was explained as simply observing and recording the maze-learning performance of the maze-bright and maze-dull rats. Half the experimenters were told that they had been assigned rats that were maze-bright while the remaining experimenters were told that they had been assigned rats that were maze-dull. None of the rats had really been bred for maze-brightness or maze-dullness, and experimenters were told purely at random what type of rats they had been assigned. Despite the fact that the only differences between the allegedly bright and dull rats were in the minds of the experimenters, those who believed their rats were brighter obtained brighter performance from their rats than did the experimenters who believed their rats were duller. Essentially the same results were obtained in a replication of this experiment employing Skinner boxes instead of mazes.

—Robert Rosenthal

See also PYGMALION EFFECT

REFERENCES

- Pfungst, O. (1965). *Clever Hans*. Translated by C. L. Rahn. New York: Holt, Rinehart and Winston. (Original work published 1911.)
- Rosenthal, R. (1976). *Experimenter effects in behavioral research: Enlarged edition*. New York: Irvington Publishers, Halsted Press Division of Wiley.
- Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart and Winston.
- Rosnow, R. L., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. New York: W. H. Freeman.

EXPERT SYSTEMS

Expert systems (ES), also called *knowledge-based systems*, are ARTIFICIAL INTELLIGENCE (AI) computer programs with sophisticated decision-making and reasoning capabilities for limited domains of knowledge. Usually using ABDUCTION (reasoning toward an inference), they are more than smart databases because they mimic human thinking. For social scientists, ES offer ways to model different theoretical frameworks and assumptions for theory testing and SIMULATION. They also provide rich applications for instruction, research design, problem solving, identification, diagnosis, and control.

Typically, ES comprise four parts: (a) the knowledge base proper, which is the body of declarative knowledge (the “stuff”) and procedural knowledge (how to use the “stuff”) about a domain; (b) the inference engine, which provides means of reasoning; (c) the knowledge-acquisition interface, which specifies modes of entry of knowledge and the forms of rules; and (d) the user interface, which can give intuitive, human-like end-user access. Problems amenable to modeling by ES are of magnitude equivalent to those solvable by a telephone consultation with a human expert.

By way of example, the rule-based ES illustrated in Table 1 constitutes a parsimonious account to predict whether pregnant women will present early or late for first prenatal care. It was constructed from FIELDNOTES of interviews with women after their first prenatal visit. It performed comparably to a MULTIPLE REGRESSION ANALYSIS of the same data and was more easily interpretable.

Table 1

Rule 1	IF Encouraged by Influential Other = yes, THEN <i>present early</i>
Rule 2	IF Encouraged by Influential Other = no, OR Encouraged by Mother = yes, THEN <i>present late</i>
Rule 3	IF Encouraged by Influential Other = indeterminate (from interview data) AND Early Symptoms = yes, THEN <i>present early</i> , ELSE (i.e., if no) <i>present late</i>

Although frequently written in AI languages such as C++, Prolog, or Lisp, many ES are developed using programming environments called

shells. A typical shell offers three of the four critical parts of an expert system—the inference engine, the knowledge-acquisition interface, and the user interface—lacking only the knowledge base provided by the developer.

Social scientists find creating an ES a useful, sobering exercise in formulating and checking their understandings because ES require specificity of both assumptions and heuristics (the informal rules of thumb that experts use in manipulating knowledge). Permitting abduction-based, INDUCTION-based, and PROBABILISTIC reasoning, an ES is more a MODEL than a rule-based approach.

Related methods include data mining, an inductive approach, and fuzzy systems, where class-inclusion may have fuzzy boundaries. Disadvantages are the inability to deal with unexpected relations and maintenance, although ES shells are more self-documenting than is, say, C++.

Benfer, Brent, and Furbee (1991) provided an accessible introduction for social scientists. Online resources are rich and include the sites www.aaai.org/AITopics/html/expert.html (with accessible general and how-to information) and www.pcai.com/ai_info/expert_systems.html (with general information, from a business perspective, and resources including freeware and shareware).

—N. Louanna Furbee

REFERENCE

Benfer, R. A., Brent, E. E., Jr., & Furbee, L. (1991). *Expert systems*. Newbury Park, CA: Sage.

EXPLAINED VARIANCE. See
R-SQUARED

EXPLANATION

A scientific explanation of an event or regularity is an argument that purports to demonstrate why the event or regularity came to pass, given other features of the world. Why was the outcome necessary or probable in the circumstances and in the presence of relevant governing laws or regularities? The most general model of a scientific explanation is that of a “covering-law

explanation.” A covering-law explanation is an argument consisting of one or more general laws, one or more particular statements (boundary conditions), and deductive derivation of a statement of the phenomenon to be explained (the explanandum). Such explanations are designed to show why the outcome was necessary in the circumstances, given the initial conditions and the laws of nature. A covering-law explanation subsumes the phenomenon under the general laws. A probabilistic explanation has a similar logic. Probabilistic explanations identify one or more statistical laws and subsume the phenomenon under these laws: Given probabilistic laws L_i , the probability of O is P. These approaches to the logic of scientific explanation give primacy to the role of scientific laws or laws of nature in explanation. Using this approach, we have explained an outcome when we have shown how it is necessary (or probable), given the relevant laws of nature.

A different approach to scientific explanation proceeds from the point of view that the world is a system of causal processes and mechanisms. Using this approach, we have explained an outcome when we have provided an account of the causal mechanisms and powers that led to the occurrence of the outcome. Causal explanations proceed by identifying causal mechanisms through which the initial conditions brought about the explanandum. The two approaches are related because the presence of causal mechanisms also implies the availability of lawlike generalizations that can function within covering-law explanations. The causal mechanism approach, however, comes closer to the intellectual task of scientific explanation. We want to know why the event occurred when we ask for an explanation of an outcome. Ptolemy’s explanation of the observed locations of the planets in the heavens proceeded on the basis of subsumption of planetary motion under a set of lawlike generalizations (cycles and epicycles). The explanation was unsatisfactory because it rested upon a false conception of the causal processes that resulted in astronomical observations (geocentric rather than heliocentric motion of the planets).

—Daniel Little

REFERENCES

- Achinstein, P. (1983). *The nature of explanation*. New York: Oxford University Press.
 Hempel, C. (1965). *Aspects of scientific explanation, and other essays in the philosophy of science*. New York: Free Press.

- Kitcher, P., & Salmon, W. C. (Eds.). (1989). *Scientific explanation* (Minnesota Studies in the Philosophy of Science, Vol. 13). Minneapolis: University of Minnesota Press.
- Little, D. (1991). *Varieties of social explanation: An introduction to the philosophy of social science*. Boulder, CO: Westview.
- Miller, R. W. (1987). *Fact and method: Explanation, confirmation and reality in the natural and the social sciences*. Princeton, NJ: Princeton University Press.
- Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

EXPLANATORY VARIABLE

An explanatory variable is a factor that is used by the data analyst to explain a phenomenon. In a REGRESSION type of model, it is also known as an INDEPENDENT VARIABLE, a regressor, an input variable, a right-hand-side variable, or an X variable.

—Tim Futing Liao

EXPLORATORY DATA ANALYSIS

Exploratory data analysis (sometimes abbreviated as EDA) consists of an approach to data analysis that allows the data themselves to reveal their underlying structure and that gives the researcher a “feel” for the data. It relies heavily on graphs and displays to reach these goals because visual inspection offers special insight into the data, but it also uses many numerical techniques. Both the visual and numerical techniques focus on searching for knowledge that allows more effective testing of theories and hypotheses. Given that exploratory data analysis represents a general philosophy of understanding data, the techniques serve as a means to that end rather than ends in themselves.

The philosophy of exploratory data analysis can be said to value two traits: openness and skepticism (Hartwig & Dearing, 1979, pp. 9–12). Analysts should be open to unanticipated data patterns that go beyond the planned MODEL and expectations. Analysts should also be skeptical of numerical summaries of data that can conceal or misrepresent the most informative aspects of the data.

HISTORICAL DEVELOPMENT

The seminal works on exploratory data analysis come from John W. Tukey (1977) and Frederick Mosteller and John W. Tukey (1977). Tukey (1977, p. vii) noted that over the 20th century, confirmatory data analysis, which assesses how precisely SAMPLE statistics can be used to make inferences about POPULATION parameters, had come to dominate statistical research. He believed that although exploratory and confirmatory data analyses emphasize different principles, both should be used by practicing researchers. In aiming to complement the dominant methods of statistical analysis, he argued that researchers cannot get along without confirmatory data analysis but need not start with it. His central principle follows from this point (Tukey, 1977): “It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it” (p. v).

Consider the differences between confirmatory and exploratory data analysis (NIST/SEMATECH, 2003):

- Confirmatory approaches begin with a prespecified model and use the parameters obtained from the data analysis to evaluate the model, whereas exploratory approaches begin with analysis of the data to infer the model and specify its ASSUMPTIONS.
- Confirmatory approaches are formal and rigorous in efforts to evaluate a model, whereas exploratory approaches are more informal and active in the effort to discover meanings contained in the data.
- Confirmatory approaches impose a model on the data, whereas exploratory approaches let the data suggest appropriate models.
- Confirmatory approaches filter the data in reducing the information to a small number of parameters, whereas exploratory approaches rely on techniques that reflect all parts of the data.
- Confirmatory approaches require strong assumptions, whereas exploratory approaches make few assumptions or attempts to validate assumptions before model testing.
- Confirmatory approaches have much power to precisely test a HYPOTHESIS and estimate parameters under narrow circumstances, whereas exploratory approaches have greater generality and less sensitivity.

The two approaches represent ideal types that, in good research, both receive attention, but Tukey and others have had considerable influence in countering a belief that mere description and exploration of data were inferior to formal hypothesis testing and model confirmation. Today, most empirical social scientific research relies on confirmatory models, such as in the use of REGRESSION and analysis of variance to test theories, but exploratory techniques have become essential parts of the research process leading to the confirmatory models.

EXAMPLES

The long volumes of Tukey (1977) and Mosteller and Tukey (1977) offer dozens of techniques and statistics—as well as a new terminology—for exploratory data analysis. A few examples of the techniques and statistics, and how they contribute to the goals of exploratory data analysis, can illustrate their potential use for researchers.

To uncover hidden data structures, the use of displays and graphs in exploratory data analysis helps researchers to notice what they never expected to see. A STEM-AND-LEAF DISPLAY and a box-and-whisker plot (see BOXPLOT) each depict the full DISTRIBUTION of the values of a single VARIABLE in a visual way, and a SCATTERPLOT depicts the relationship between two variables with all values of the variables. The visual representations do much to identify any DEVIATION from normality (in a single variable) and from linearity (in the relationship between two variables), and thereby help evaluate assumptions commonly made in the use of LEAST SQUARES regression.

To counter non-normality and NONLINEARITY, exploratory data analysis searches for ways to reexpress the SCALE of measurement of variables. Reexpression alters the distance between values of a variable while maintaining the sequence or ordering. Taking powers, roots, and logs of variables normalizes distributions and straightens relationships, making the use of the transformed variables more appropriate for multivariate analyses. Techniques to smooth sequential variables similarly help to reveal underlying patterns in the data.

To best summarize distributions and relationships, exploratory data analysis relies on the computation of ROBUST and resistant measures. By reflecting the bulk of the data, resistant measures are less sensitive to DEVIANT CASE ANALYSES. Such measures are often

based on a MEDIAN rather than a mean, on absolute values rather than squared values, and on least absolute deviations rather than least squared deviations. Various robust regression techniques further provide ways to make models less sensitive and discount the disproportional influence of outlying cases.

To evaluate the suitability of a model, exploratory data analysis gives much attention to OUTLIERS and RESIDUALS. Least squares analyses of relationships among variables are accompanied by visual inspections of the scatterplots and deviations of points from the regression line. Many numerical techniques also identify large outliers and residuals in a multivariate context but cannot fully replace the benefits obtained from the visual contrast of the model with the data. In turn, the attention to outliers and residuals helps determine how to reexpress variables, respecify models to better describe relationships, and gain new insights into the topics under investigation.

—Fred C. Pampel

REFERENCES

- Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis* (Sage University Papers on Quantitative Applications in the Social Sciences). Beverly Hills, CA: Sage.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.
- NIST/SEMATECH. (2003). *E-handbook of statistical methods*. Retrieved from <http://www.itl.nist.gov/div898/handbook/eda>
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

EXPLORATORY FACTOR ANALYSIS.

See FACTOR ANALYSIS

EXTERNAL VALIDITY

External validity is a property that allows research findings to be generalized to a larger POPULATION. External validity is a concern when constructing experimental and non-experimental (OBSERVATIONAL) RESEARCH designs. Researchers can ensure external validity through careful construction of their research design.

There are three main threats to external validity: nonrepresentative SAMPLES, an artificial laboratory environment, and testing effects. One of the main goals of research is to generalize the findings to a larger population. Consider a study on the effect that negative campaign advertising has on voters' approval of the incumbent candidate. For the study to be feasible, it cannot be administered to all possible voters, the population of interest. Instead, a sample of the population is selected to take part in the study. Special attention needs to be paid so that the underlying causal process is the same for both the sample and the population of interest. Finding a REPRESENTATIVE SAMPLE is sometimes difficult. Many times, participants are recruited from local communities or college campuses where the study is taking place. A campaign advertising study based on a sample of college juniors is unlikely to resemble the results that would come from the population of interest. College juniors may be more easily swayed by the advertisements because they have not formed strong partisan ties, whereas older voters may be less affected by the negative advertisements.

The setting in which a study is carried out can have an impact on the findings. Experiments tend to study phenomena in a laboratory setting. This unnatural environment, where participants know they are being studied, may produce unintended results. In the negative advertising study, participants may be asked to view various advertisements, and then their opinion of the incumbent candidate would be measured. However, we know most people do not actively pay attention to commercials on television; rather, during commercial breaks, many people hold conversations, get snacks, or change stations. Thus, the artificial environment in which the advertisements are viewed could cause participants to evaluate the incumbent poorly, though the same commercials viewed at home could have no effect. Observational studies may encounter problems with artificial settings if the observation takes place in the laboratory setting versus the natural environment.

The final threat to external validity is testing. PRETESTS and the experience of being tested may change the magnitude of the treatment effect. To be able to measure the effect of negative advertising on incumbent approval, we would need some measure of approval before the participants viewed the advertisements. A pretest could make the participants become more observant or opinionated about the incumbent and produce misleading results. In observational studies, the researcher must ensure that the presence

of the observer will not influence the behavior of the subjects.

Researchers increase the external validity of their studies by being mindful of the threats when creating their research design. Iyengar, Peters, and Kinder (1982) illustrated how this can be done in an exploration of the agenda-setting power of the evening news on television. They manipulated the volume of stories shown in an evening news broadcast and observed whether it had an impact on how participants viewed the issues.

Iyengar, Peters, and Kinder's sample was drawn from volunteers from New Haven, Connecticut. This could pose a problem because the population of New Haven is not representative of the population of the United States (the population of interest). To ensure that their results would be generalizable to the population of interest, they attempted to encourage diversity by soliciting participation through a classified advertisement. Increasing heterogeneity of the sample can make the research more generalizable (Frankfort-Nachmias & Nachmias, 2000). They also reported on the demographic makeup of the participants, allowing the reader to make judgments about the generalizability of the experiment. They also used RANDOM ASSIGNMENT between their control and experimental groups to ensure the absence of systematic differences between these groups.

Second, the researchers took several steps to prevent artificial findings from being generated by the process of experimentation. The participants watched the news broadcasts in small groups and were never encouraged to pay particular attention to the broadcasts. In fact, the researchers reported that many of the participants engaged in conversations during the broadcasts. This indicates that although the experiment did not take place in the participants' living rooms (the environment of interest), the researchers were successful in reproducing a similar environment, including distractions. Careful consideration should be paid to the location of the testing environment to make it resemble the natural setting as much as possible. Another option is to use a randomized FIELD EXPERIMENT (Gerber & Green, 2000).

The researchers also took several precautions to prevent spurious results from testing. Some of these included presenting a convincing COVER STORY to participants and designing the pretest in a nonleading way. A cover story may be a useful tool; however, researchers also need to be concerned with

the ethical implications of using a cover story to deceive participants. It is through careful consideration of the threats to external validity that Iyengar et al. were able to construct a research design that minimized the impact of these threats.

Frankfort-Nachmias and Nachmias (2000) explained that researchers need to be conscious of the trade-off between external validity and INTERNAL VALIDITY when constructing a research design. Internal validity ensures that the INDEPENDENT VARIABLES are producing the change in the DEPENDENT VARIABLE. An experiment with strong external validity but weak internal validity contributes very little to advancing our understanding because conclusions cannot be drawn about the RELATIONSHIP being studied. To ensure internal validity, many times we need to sacrifice external validity. In Iyengar et al.'s (1982) experiment on agenda setting, having the participants watch the news in their living rooms would have provided a more natural setting. However, this would have prevented the researchers from being able to control what the participants were viewing. Thus, to ensure internal validity, they sacrificed some external validity by holding the experiment in a controlled environment. When constructing a research design, one needs to be aware of the impact that choices have on both external and internal validity.

—Heather L. Ondercin

REFERENCES

- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Frankfort-Nachmias, C., & Nachmias, D. (2000). *Research methods in the social sciences* (6th ed.). New York: Worth.
- Gerber, A. S., & Green, D. P. (2000). The effects of personal canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *The American Political Science Review*, 94, 653–664.
- Iyengar, S., Peters, M. E., & Kinder, D. R. (1982). Experimental demonstrations of the “not-so-minimal” consequences of television news programs. *The American Journal of Political Science*, 4, 848–858.

EXTRAPOLATION

Also known as *trend extrapolation* and *curve fitting*, extrapolation is a projection technique in which

the analyst plots data points from past time periods, chooses a best-fitting trend line (or curve) for this data, and then extends that trend line to project future values. At its core, extrapolation is a bivariate REGRESSION technique in which the INDEPENDENT VARIABLE is time and the variable to be projected is the DEPENDENT VARIABLE.

Smith, Tayman, and Swanson (2001) noted that the defining characteristic of the extrapolation technique is that historical values are used exclusively to project future values of the variable being studied. There are numerous variations of the extrapolation technique, ranging from simple linear regression to more complex polynomial and logistic curves, with some variations also incorporating ratios into the projection model.

Extrapolation is employed routinely by demographers and planners to project population figures, by public administrators to project budget revenues and expenditures, and by businesspersons to project sales and revenues. Extrapolation's popularity rests primarily in its simplicity and its low data requirements. Using a spreadsheet program or calculator, projections using the extrapolation technique can be generated easily. As for data requirements, as few as two data points are needed to provide a trend line to project future values. Although extrapolation is relatively simple in application, Smith et al. (2001) reported that numerous empirical studies have shown that it is as accurate as other, more complex methods, such as the cohort-component technique.

Figure 1 shows the results when extrapolation is used to project the population of a medium-sized county using three variations: a linear curve, a geometric curve, and a parabolic curve. Using data from the U.S. Census Bureau, total population figures from 1940 to 2000 were used to project the county's population for 2010, 2020, and 2030. Figure 1 shows that population projections range markedly based on the curve chosen. At face value, the parabolic method appears to be the “best fitting” curve for this situation.

The simplicity and low data requirements that make extrapolation attractive as a projection method also serve as the primary limitations to the technique. Extrapolation relies upon aggregated data inputs and yields aggregated results. Factors that may influence the variable under study remain external to the method. For example, when undertaking population projections, the extrapolation technique does not consider the effects of housing trends, economic changes, or other external pressures on population. The inputs into

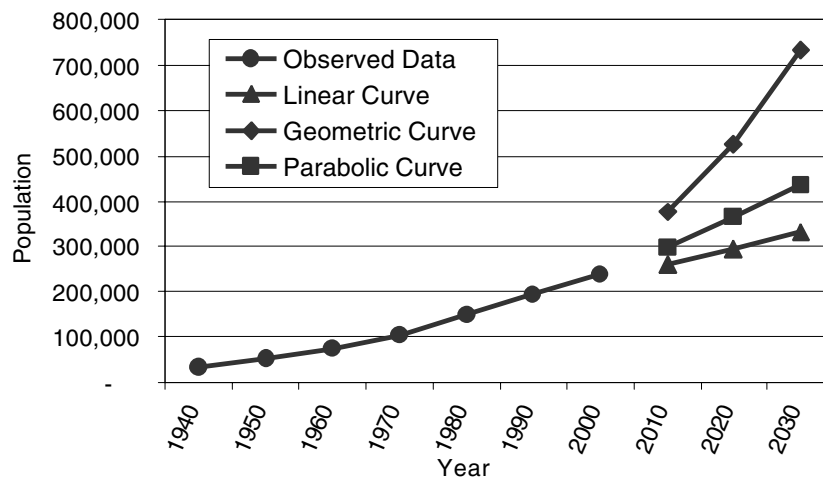


Figure 1 Example Application of the Extrapolation Technique

the model are simple, and so are the outputs. When using extrapolation to project government revenues, for example, only a total revenue figure is generated by the technique; no detail is provided on changes in revenues across different funds.

A further limitation to the model is that extrapolation uses past conditions to project future conditions. Although past conditions are a useful starting point for predicting the future, there is no assurance that past trends will continue. Last, the choice of time period and the number of data observations employed in the analysis can greatly influence the results. For these reasons, extrapolation should be employed carefully, with full understanding of its utility and limitations for projecting the future.

—Tim Chapin

REFERENCE

Smith, S. K., Tayman, J., & Swanson, D. A. (2001). *State and local population projections: Methodology and analysis*. New York: Kluwer Academic/Plenum Publishers.

EXTREME CASE

An extreme case is a CASE that has a value on a variable that is at the limit of the possible values. For example, in the study of age in a community, the most extreme high value may be 95 years, and the most extreme low value may be 18 years.

These represent extreme cases, but they may or may not be OUTLIERS. An outlier necessarily falls far from the other values, but an extreme case may not do this. For example, in an age study, there may be many respondents in their 80s and 90s, which would mean that an age of 95, though extreme, does not really lie out from neighboring values. In contrast, if below 95, the next highest age was 72, then clearly 95 would be an outlier, as well as being an extreme case.

—Michael S. Lewis-Beck

F

***F* DISTRIBUTION**

The *F* distribution, also called the variance-ratio distribution, is the SAMPLING DISTRIBUTION derived from the ratio of two sample variances (s^2) estimating identical population VARIANCES (σ^2). The distribution was originally proposed by Sir Ronald A. Fisher (1890–1962) and was developed by George W. Snedecor (1881–1974), who named it the *F* distribution in honor of Fisher.

SIGNIFICANCE TESTING WITH THE *F* DISTRIBUTION

Consistent with its development by Fisher and Snedecor, the *F* distribution is most often used in conjunction with the ANALYSIS OF VARIANCE. However, in its most generic SIGNIFICANCE TESTING application, the *F* distribution is useful for testing hypotheses about the equivalence of two population variances:

$$\begin{aligned}H_0 : \sigma_1^2 &= \sigma_2^2 \\ H_1 : \sigma_1^2 &\neq \sigma_2^2.\end{aligned}$$

To understand the *F* distribution, and its utility in testing such hypotheses, let's first imagine that H_0 is true. That is, consider two normally distributed populations with means (μ) that may differ, but with identical variances (σ^2). If we were to take a random sample of size n_1 from one population and a random sample of size n_2 from the other population, we could then

estimate the population variance from each sample by constructing a sample variance:

$$\hat{\sigma}^2 = s^2 = \frac{\sum(X - \bar{X})^2}{n - 1}.$$

The denominator of the sample variance ($n - 1$) is the DEGREES OF FREEDOM associated with that statistic, which we can label ν . Even though the two populations have identical variances, SAMPLING ERROR would likely lead to different sample variances. We could then construct a ratio of the two sample variances, which would be the *F* RATIO:

$$F = \frac{s_1^2}{s_2^2} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}.$$

With H_0 true, we could derive the sampling distribution of the *F* ratio by constructing the *F* ratio for every possible combination of samples of size n_1 and n_2 , which results in the *F* distribution for ν_1 (i.e., $n_1 - 1$) and ν_2 (i.e., $n_2 - 1$). (With H_0 false, the distribution—called the noncentral *F* distribution—is not easily specified, but it would be more symmetrical than the *F* distribution.)

As an example of how the *F* distribution could be used to test a hypothesis about the equivalence of population variances, consider two samples of $n_1 = 31$ ($\nu_1 = 30$) and $n_2 = 41$ ($\nu_2 = 40$). Because we are most often interested in the tails of a distribution, wherein reside atypical values, statistics texts usually contain a table of *F* ratios that determine the upper percentages of the *F* distribution (upper 2.5%, upper 5%, etc.). If one had adopted an α level of .05, the *F* ratio that would cut off the upper 5% of the

F distribution with $\nu_1 = 30$ and $\nu_2 = 40$ (called F_{Critical}) would be 1.74. Thus, according to the F distribution, only 5% of the F ratios would be 1.74 or greater when H_0 is true. (Similarly, statistical programs will generate a probability value that the F ratio would have been obtained with H_0 true.) Armed with such information, we would be able to test null hypotheses that two sample variances were drawn from populations with identical variances. If the F ratio for our two samples were 1.74 or greater (or the probability less than our α level, typically .05), we would be inclined to think that the numerator sample had been drawn from a population with a greater variance than that from which the denominator sample had been drawn. If the F ratio is greater than 1.0 but less than 1.74, we would retain H_0 . (If the F ratio is less than one, simply invert the ratio and compare it to F_{Critical} with 40 and 30 degrees of freedom, which would be 1.79 with $\alpha = .05$.)

CHARACTERISTICS OF THE F DISTRIBUTION

Given that the shape of the F distribution is driven by the degrees of freedom (ν_1 and ν_2) associated with the two sample variances, there is not a single F distribution, but a family of distributions. The F ratios in the F distribution will always fall between zero and infinity. As long as ν_2 is greater than 2, the mode of the F distribution would be

$$\left(\frac{\nu_1 - 2}{\nu_1} \right) \left(\frac{\nu_2}{\nu_2 + 2} \right).$$

As long as ν_2 is greater than 2, the mean of the F distribution would be

$$\frac{\nu_2}{\nu_2 + 2}.$$

As long as ν_2 is greater than 4, the variance of the F distribution would be

$$\frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}.$$

The F distribution typically will be unimodal and positively skewed (as long as ν_1 is greater than 2). In a positively skewed distribution, the mean would be larger than the mode. For the example above, with $\nu_1 = 30$ and $\nu_2 = 40$, the modal F ratio would be .89 and the mean F ratio would be .95, consistent with the positive skew of the F distribution. From the formulas, it should be clear that the mode will always be less

than the mean, although the difference will lessen as ν_1 approaches infinity.

The F distribution is related to another commonly used statistical distribution, the T-DISTRIBUTION, such that when $\nu_1 = 1$, $t^2 = F$.

—Hugh J. Foley

See also DISTRIBUTION

REFERENCES

- Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.
- Heyde, C. C., & Seneta, E. (Eds.). (2001). *Statisticians of the centuries*. New York: Springer.
- Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice Hall.
- Stuart, A., & Ord, J. K. (1994). *Kendall's advanced theory of statistics* (Vol. 1). London: Edward Arnold.

F RATIO

The F ratio is the ratio of variances or variance-like quantities. A simple example is the testing of whether the variance between groups is significantly larger than the variance within groups in ANALYSIS OF VARIANCE. The general notion underlying such a test is that the differences between individuals within groups are only random differences (error variance), but that the differences between group means stand out above chance (alternative hypothesis).

AN EXAMPLE

An example of calculating this can be found when the partition of variance is dealt with. A series of scores 2 2 3 4 4 5 6 7 8 9 with a mean of 5 and variation 54 is partitioned into two groups. We can think of two groups (e.g., a group of men and a group of women) for which the scores of a quantitative variable are considered—let us say, their smoking behavior measured as the number of cigarettes per day. The group of men with scores 2 2 3 4 4 has a mean of 3 and a variation of 4. The group of women with scores 5 6 7 8 9 has a mean of 7 and a variation of 10. The within-group variation is $4 + 10 = 14$. The between-group variation is defined by the group means 3 and 7. Their mean is 5 and their variation 8. The between-group variation entirely determined by the group is obtained by multiplying the variation between groups by the number of individuals

per group: $8 \times 5 = 40$. It appears that $54 = 40 + 14$ or $SST = SSB + SSW$ (sum of squares total = sum of squares between + sum of squares within). An analogous partition of the total into a between and within component can be carried out for the degrees of freedom. One degree of freedom is lost to the mean of a series. Therefore, there are 9 degrees of freedom for the total of 10 scores. For the between component, we calculate for 2 groups, so that there is 1 degree of freedom. For the within component, we calculate for 2×5 individuals, each time losing a degree of freedom to the group mean, so that $2(5 - 1) = 8$ remain. We see here, too, that total = between + within, because $9 = 1 + 8$. Should one now want to work with variances instead of variations (Mean squares, MS, instead of Sum of Squares, SS), then one must divide each of the variations concerned by the corresponding degrees of freedom: $MSB = 40/1$ and $MSW = 14/8$. The ratio of this between-variance MSB and the within-variance MSW is a statistical quantity $F = (40/1)/(14/8) = 22.86$, for which the sampling distribution for the degrees of freedom, 1 for the numerator and 8 for the denominator, is distributed as F (the symbol F was chosen to commemorate the contributions by Sir Ronald Fisher). The critical value of F for 1 degree of freedom for the numerator and 8 degrees of freedom for the denominator is found in statistical tables of the theoretical F distribution and is equal to 5.32 for a postulated α of .05. The value of the F ratio we found in our empirical example is greater, and we can therefore, even with these small numbers, reject the null hypothesis with a probability of 95%. The null hypothesis states that the ratio MSB/MSW is equal to 1; in other words, that the differences between groups (MSB) are not greater than the randomly assumed differences between individuals within groups (MSW = error variance).

MANY APPLICATIONS IN MULTIVARIATE ANALYSIS

In multivariate analysis, the F ratio is frequently used in situations more complicated than the example above. The test of difference of groups in analysis of variance, the test for the significance of the regression model in multiple REGRESSION analysis and the test for centroid differences in DISCRIMINANT ANALYSIS are just a few of the many examples of application. C. R. Rao was especially charmed by the F distribution and has developed an F variant for many applications in multivariate analysis.

RELATIONSHIP BETWEEN F AND OTHER DISTRIBUTIONS

The attentive reader will already have seen that instead of the F test for this example, we could just as easily have used the t test of difference of means. We could have situated the obtained difference (-4) between the mean of the first group (3) and the mean of the second group (7) in a sampling distribution of differences of means with the estimated standard error $= (\frac{s_w^2}{n_1} + \frac{s_w^2}{n_2})^{1/2}$ and where s_w^2 is equal to $(4 + 10)/[(5 - 1) + (5 - 1)] = 1.75$ and $n_1 = n_2 = 5$. The t value for the difference of means found in our pair of samples would be equal to $t = (4 - 0)/[(1.75/5) + (1.75/5)]^{1/2} = 4.78$. From the t table, a significant difference would then appear between the means of the two groups. The F value of 22.86 obtained earlier is (within rounding errors) equal to the square of the t value found of 4.78, because when the degree of freedom is one, $F = t^2$.

One can also see in the F and t tables that the critical values of F for a numerator of 1 degree of freedom (only two groups!) are equal to the squares of the corresponding values of t .

There are close connections between the F DISTRIBUTION and many other statistical DISTRIBUTIONS. The F distribution belongs to a subfamily of beta distributions, sometimes called *inverted beta distributions*. There is also a connection with the t , the chi square, the NORMAL DISTRIBUTION, and others. See F DISTRIBUTION for further discussions.

—Jacques Tacq

REFERENCES

- Blalock, H. M. (1972). *Social statistics*. Tokyo: McGraw-Hill Kogakusha.
- Fisher, R. A. (1953). *The design of experiments*. Edinburgh, UK: Oliver and Boyd.
- Hays, W. L. (1972). *Statistics for the social sciences*. New York: Holt International.
- Tacq, J. J. A. (1997). *Multivariate analysis techniques in social science research: From problem to analysis*. London: Sage.

FACE VALIDITY

Face validity is an estimate of the degree to which a measure is clearly and unambiguously tapping the construct it purports to assess. Thus, face validity refers

to the “obviousness” of a test—the degree to which the purpose of the test is apparent to those taking it. Tests wherein the purpose is clear, even to naïve respondents, are said to have *high face validity*; tests wherein the purpose is unclear have *low face validity* (Nevo, 1985). The concept of face validity is similar to *item subtlety*, but there are important differences as well. Whereas face validity describes the transparency of an entire test, item subtlety describes the transparency of individual test items (Bornstein, Rossner, Hill, & Stepanian, 1994). It is possible to construct a test wherein the purpose of individual test items is not apparent, but when these items are scrutinized as a group, the purpose of the test as a whole becomes obvious.

Face validity has contrasting effects on different types of tests. Studies indicate that high face validity can facilitate performance on intelligence, aptitude, and achievement tests: When the purpose of the test seems clear, testees are less anxious and more motivated to persevere, even when test items are highly challenging (Messick, 1995; Nevo, 1985).

High face validity can be a liability when a test is designed to assess some aspect of personality or psychopathology. In this situation, high face validity enables respondents to bias their responses to present themselves as they want to be seen by the examiner. Studies show that naïve respondents are able to “fake good” and “fake bad” more effectively on tests with high face validity than those with low face validity (Bornstein, 2002; Bornstein et al., 1994).

There are two general methods for evaluating the face validity of a test. Some psychometricians have taken a direct approach, asking participants to identify the purpose of an assessment instrument from among an array of likely choices. To the degree that participants can do this accurately, the test has high face validity. Other researchers have taken an indirect approach, asking participants to deliberately bias their answers to raise or lower their scores on the test. To the degree that participants can alter their scores in this way, the test is presumed to have high face validity and to be susceptible to self-presentation effects in vivo (Bornstein, 2002).

Given its importance for psychological tests, face validity should always be assessed and controlled during test development. Several strategies are useful. For example, test items can be worded subtly, so the true purpose of the measure is masked. Alternatively, test items can be ordered so that those from a given subscale do not appear in proximity to each other; this

has also been shown to lower the face validity of the test. Finally, distracter items unrelated to the true purpose of the test can be included so that the content of genuine test items is less obvious. This latter strategy is particularly effective, when combined with the other two approaches, for disguising the purpose of the test.

—Robert F. Bornstein

REFERENCES

- Bornstein, R. F. (2002). A process dissociation approach to objective-projective test score interrelationships. *Journal of Personality Assessment*, 78, 47–68.
- Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L. (1994). Face validity and fakability of objective and projective measures of dependency. *Journal of Personality Assessment*, 63, 363–386.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.
- Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, 22, 287–293.

FACTOR ANALYSIS

In general, the term *factor analysis* refers to any one of a number of similar but distinct multivariate statistical models that model observed variables as linear functions of a set of LATENT or hypothetical variables that are not directly observed, known as *factors*.

Factor analysis models are similar to REGRESSION models in that they possess DEPENDENT VARIABLES that are linear functions of INDEPENDENT VARIABLES. But unlike regression, the independent variables of the factor analysis models are not observed independently of the observed dependent variables.

Factor analysis models may be further distinguished according to whether the factor variables are determinate or not. Factors are determinate if they can be derived in turn as linear functions of the observed variables. Otherwise, they are indeterminate. Determinate models encompass the various component analysis models, such as PRINCIPAL COMPONENTS ANALYSIS (Hotelling, 1933; Joliffe, 1986; Pearson, 1901); weighted principal components (Mulaik, 1972); and Guttman's image analysis (Guttman, 1953). Indeterminate models are represented by the common factor model (Spearman, 1904; Thurstone, 1947),

which seeks to account for the covariation between the observed variables as the result of the observed variables' sharing in varying degrees the influences of the variation of a common set of common factor variables.

Determinate factor analysis models are often useful in a data reduction role by finding a smaller number of variables that capture most of the information of variation and covariation among the observed variables. Scores on the determinate factors can be computed as linear combinations of the observed variables. These factor scores may be used as independent and dependent variables—as the case warrants—in other multivariate statistical procedures, such as multivariate regression or multivariate analysis of variance. However, for substantive theoretical work, the main drawback of these determinate component analysis models is that their factors represent statistical artifacts unique to the set of observed variables determining them (Mulaik, 1987). Change the set of observed variables, and you obtain different linear combinations. The component factors have no independent existence apart from the set of observed variables of which they are linear combinations.

In contrast, the common factors of the common factor model are indeterminate from the observed variables and are not linear combinations of them. Common factors can correspond to variables having an independent existence, and, in the theory of simple structure in common factor analysis (Thurstone, 1947), different sets of observed variables from a domain can be linear functions of the same common factors. Thus, for the purposes of discovering autonomous variables that have theoretical import as common causes of other variables, the common factor model is generally preferred to a determinate component analysis model.

However, the common factor model has limits to its application. Common factor analysis is limited to the case where there is no natural order among the observed variables. The only ordering principle permitted by the common factor model is the relation of functional dependency of observed variables on latent variables. No provision exists in the common factor model for functional dependencies between latent variables. Relations between latent variables only take the form of correlations or covariances, which are nondirectional and nonordering. But variables that are naturally ordered in time, space, or degree of some attribute may be more correlated with the variables immediately adjacent to them and have diminishing correlations

with variables farther from them in the natural ordering. In these cases, in addition to functional dependencies of observed variables on latent variables, each successive latent variable corresponding to one of the observed variables in the order may be a linear function of only the immediately preceding latent variable in the order plus some new latent variable unique to it. Models with this property are known as simplex models (Jöreskog, 1979). To deal with this and other cases, STRUCTURAL EQUATION MODELING is more appropriate because it has provisions for establishing linear functional relations between latent variables. Thus, if one is to apply the common factor model properly to a set of observed variables, one must take care to determine at the outset that there is no apparent natural ordering among the variables. Although the common factor model may fit well to such data with a small number of common factors, the factors will usually defy theoretical interpretation (Jones, 1959).

Factor analysis models also can be distinguished as to whether they are exploratory or confirmatory. Exploratory common factor analysis is principally analytic, used in situations where there is minimal knowledge of the constituent factor variables of a domain of variables believed to conform to the common factor model. The aim is to analyze the observed variables to “discover” constituent variables that are more basic. Confirmatory factor analysis is principally synthetic. It is used where there is sufficient knowledge to formulate hypotheses as to how certain theoretical variables function as factors common to a number of observed variables. The theory allows one to predict within certain constraints the pattern of covariances among the observed variables and to test for this pattern (see CONFIRMATORY FACTOR ANALYSIS). The rest of this article will be about exploratory common factor analysis.

EQUATIONS OF EXPLORATORY FACTOR ANALYSIS

The equations of factor analysis are usually expressed in matrix algebra. Thus, the model equation of common factor analysis is given by the following matrix equation:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E},$$

where $\mathbf{Y}$ is a $p \times 1$ column vector of p observed random variables; $\mathbf{A}$ is a $p \times m$ matrix of factor pattern

loadings λ_{ij} , which are analogous to regression coefficients; $\mathbf{X}$ is an $m \times 1$ column vector of common factor random variables; and $\mathbf{E}$ is a $p \times 1$ column vector of p unique factor random variables.

One then assumes the following:

(a) The common and unique factor variables are uncorrelated, that is, $\text{cov}(\mathbf{X}, \mathbf{E}) = 0$; and (b) the unique factors are mutually uncorrelated, that is, $\text{cov}(\mathbf{E}) = \mathbf{\Theta}^2 = \text{a diagonal matrix}$.

From the model equation and these assumptions, we are able to derive the fundamental theorem of the common factor model, expressed as

$$\text{cov}(\mathbf{Y}) = \mathbf{\Sigma} = \mathbf{\Lambda} \mathbf{\Phi} \mathbf{\Lambda}' + \mathbf{\Theta}^2,$$

where $\mathbf{\Sigma}$ is the $p \times p$ variance-covariance matrix for the p observed variables, $\mathbf{\Lambda}$ is the $p \times m$ factor pattern matrix of elements λ_{ij} , $\mathbf{\Phi}$ is the $m \times m$ matrix of VARIANCES and COVARIANCES among the common factors, and $\mathbf{\Theta}^2$ is a $p \times p$ diagonal matrix of unique factor variances. Because $\mathbf{\Theta}^2$ is a diagonal matrix with zero off-diagonal elements, the off-diagonal elements of the covariance matrix $\mathbf{\Sigma}$ are due only to the off-diagonal elements of $\mathbf{\Lambda} \mathbf{\Phi} \mathbf{\Lambda}'$, which is a function of only the common factor variables.

STEPS IN PERFORMING A FACTOR ANALYSIS: A WORKED EXAMPLE

Step 1: Choosing or Constructing Variables

It is best to select or construct variables representing some domain of activity in a systematic manner. L. L. Thurstone (1947), a pioneer in the field of factor analysis, believed that a domain of variables is defined by the common factors that span it. Thus, one must consider the possible common factors that might exist in the domain. Then, for each of those anticipated factors, one should select or construct at least four indicator variables that one believes are relatively pure measures of those factors. Four indicators of an anticipated factor overdetermine the factor.

As an example, Carlson and Mulaik (1993) selected 15 variables based on 15 bipolar personality rating scales shown in Table 1 that they expected would

Table 1 Fifteen Bipolar Personality Rating Scales Used to Produce 15 Variables for the Factor Analysis Study

1.	FRIENDLY:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	UNFRIENDLY
2.	SYMPATHETIC:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	UNSYMPATHETIC
3.	KIND:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	CRUEL
4.	AFFECTIONATE:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	UNAFFECTIONATE
5.	INTELLIGENT:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	UNINTELLIGENT
6.	CAPABLE:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	INCAPABLE
7.	COMPETENT:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	INCOMPETENT
8.	SMART:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	STUPID
9.	TALKATIVE:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	UNTALKATIVE
10.	OUTGOING:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	WITHDRAWN
11.	GREGARIOUS:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	SOLITARY
12.	EXTRAVERTED:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	INTROVERTED
13.	HELPFUL:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	UNHELPFUL
14.	COOPERATIVE:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	UNCOOPERATIVE
15.	SOCIABLE:	1	:	2	:	3	:	4	:	5	:	6	:	7	:	UNSOCIABLE

represent four factors of personality: (a) friendliness, represented by the bipolar rating scales friendly-unfriendly, sympathetic-unsympathetic, kind-unkind, and affectionate-unaffectionate; (b) ability, represented by the scales intelligent-unintelligent, capable-incapable, competent-incompetent, and smart-stupid; (c) extraversion, represented by talkative-untalkative, outgoing-withdrawn, gregarious-solitary, extraverted-introverted, and sociable-unsociable. They anticipated that variables such as helpful-cooperative and cooperative-uncooperative would be concepts representing a combination of friendliness and ability, and would have “loadings” on these factors plus loadings on (d) an additional factor unique to them alone.

Step 2: Obtaining Scores on Variables and Computing Correlation Matrix

Carlson and Mulaik (1993) then had 280 students rate randomly selected descriptions of people in a work setting on the 15 rating scales. The correlations among the 15 scales were then obtained and are given in Table 2.

Step 3: Determining Number of Factors

The extraction of factors is based on the matrix of correlations between factors. However, a decision must be made at this point as to which method of factor extraction to use to find the estimates for the factor pattern coefficients, the correlations among the common factors, and the unique factor variances, which will be the basis for the ultimate interpretation of the analysis. The program used to perform this analysis was SPSS FACTOR Version 4.0 for the Macintosh

Table 2 Correlations Among 15 Variables

	1 FRIENDLY																											
1 FRIENDLY	1.000	2 SYMPATHETIC																										
2 SYMPATHETIC	.777	1.000	3 KIND																									
3 KIND	.809	.869	1.000	4 AFFECTIONATE																								
4 AFFECTIONATE	.745	.833	.835	1.000	5 INTELLIGENT										SYMMETRIC MATRIX													
5 INTELLIGENT	.176	.123	.123	.112	1.000	6 CAPABLE									LOWER HALF SHOWN													
6 CAPABLE	.234	.159	.205	.183	.791	1.000	7 COMPETENT																					
7 COMPETENT	.243	.155	.187	.186	.815	.865	1.000	8 SMART																				
8 SMART	.234	.190	.238	.215	.818	.841	.815	1.000	9 TALKATIVE																			
9 TALKATIVE	.433	.319	.321	.435	.174	.209	.239	.258	1.000	10 OUTGOING																		
10 OUTGOING	.473	.480	.410	.527	.220	.274	.269	.261	.744	1.000	11 GREGARIOUS																	
11 GREGARIOUS	.433	.438	.406	.526	.188	.227	.242	.228	.711	.853	1.000	12 EXTRAVERTED																
12 EXTRAVERTED	.447	.396	.350	.500	.192	.221	.227	.224	.758	.846	.801	1.000	13 HELPFUL															
13 HELPFUL	.649	.693	.697	.694	.283	.344	.370	.365	.443	.552	.514	.473	1.000	14 COOPERATIVE														
14 COOPERATIVE	.662	.692	.676	.679	.311	.345	.375	.351	.431	.557	.514	.493	.740	1.000	15 SOCIABLE													
15 SOCIABLE	.558	.543	.510	.632	.213	.289	.287	.287	.745	.886	.820	.830	.631	.626	1.000													

SYMMETRIC MATRIX
LOWER HALF SHOWN

computer. (Later versions are essentially unchanged in these options.) The program offers several methods of factor extraction appropriate for a common factor analysis. These are principal axis factoring, which is essentially unweighted least squares; unweighted least squares (a slightly different ALGORITHM for the same result); GENERALIZED LEAST SQUARES; and MAXIMUM LIKELIHOOD. All of these methods of extraction are iterative and require that an initial decision be made as to the number of factors to extract, because this number must be fixed at this value throughout the iterations. SPSS computes a principal components analysis with every factor analysis and displays the EIGENVALUES of the unmodified CORRELATION matrix R . These can be used to estimate the number of common factors. This is done by means of a SCREE PLOT, which is a plot of the magnitude of each of the eigenvalues against its ordinal position in the descending series of eigenvalues. Connecting the dots in this plot reveals a large first eigenvalue much higher than the rest, followed by a number of lesser eigenvalues of still substantial magnitude. But there is usually a point where the rapid descent in magnitude of the eigenvalues suddenly changes to a gradual, almost linear descent for the remainder of the eigenvalues. Many believe that this is in the vicinity of the point where the variables begin to be influenced by common factors in a significant way, and so this number is used as the number of common factors to extract. Others retain only so many factors as there are eigenvalues of R greater than 1.00. But this is well known to be a weakest lowest bound for the number of common factors. There may be more of lesser influence.

This and subsequent versions of SPSS FACTOR do not report a different set of eigenvalues that would be more revealing theoretically as to the number of common factors. Given the equation of the fundamental theorem of factor analysis above, we may substitute $\mathbf{R}$ for $\mathbf{\Sigma}$ to represent a correlation matrix. Guttman (1954, 1956) showed that a strong lower bound to the number of common factors would be the number of positive eigenvalues of the matrix $\mathbf{R} - \mathbf{S}^2$ where $\mathbf{S}^2 = [\text{diag}(\mathbf{R}^{-1})]^{-1}$. $\mathbf{S}^2$ is a strong upper-bound approximation to $\mathbf{\Theta}^2$. This number of positive eigenvalues, however, is often on the order of $p/2$. Furthermore, many of the eigenvalues of $\mathbf{R} - \mathbf{S}^2$ are quite small, near zero, and so many researchers would like to examine a plot of these eigenvalues to determine a point at which the eigenvalues begin to assume substantial values. Developments by Harris (1962) that were adopted by Jöreskog (1967) in developing an algorithm for performing maximum likelihood factor analysis suggested that one could pre- and postmultiply $\mathbf{R} - \mathbf{S}^2$ by $\mathbf{S}^{-1}$ to obtain $\mathbf{S}^{-1}(\mathbf{R} - \mathbf{S}^2)\mathbf{S}^{-1} = \mathbf{S}^{-1}\mathbf{R}\mathbf{S}^{-1} - \mathbf{I}$. This transformation does not change the number of positive eigenvalues of the resulting matrix, so the number of positive eigenvalues of this matrix would equal the number of factors to retain. But each of the eigenvalues of this matrix would correspond to an eigenvalue of the matrix $\mathbf{S}^{-1}\mathbf{R}\mathbf{S}^{-1}$ minus 1. So, any eigenvalue greater than 1.0 of $\mathbf{S}^{-1}\mathbf{R}\mathbf{S}^{-1}$ would correspond to a positive eigenvalue of $\mathbf{S}^{-1}\mathbf{R}\mathbf{S}^{-1} - \mathbf{I}$. Thus, researchers could also profit from examining the eigenvalues of $\mathbf{S}^{-1}\mathbf{R}\mathbf{S}^{-1}$ to determine the number of common factors to retain.

A separate calculation in this example of the eigenvalues of $\mathbf{S}^{-1}\mathbf{R}\mathbf{S}^{-1}$ revealed only 7 of the

15 eigenvalues were greater than 1.00. These were 37.484, 14.467, 9.531, 1.436, 1.289, 1.116, and 1.065. Although three factors would be substantial in contribution, we will take one more because that was the theoretical expectation, and those beyond it would have a much smaller contribution (if you subtract one from each).

Step 4: Extraction of the Unrotated Pattern Matrix

The method of factor extraction used in this example was maximum likelihood. It is relatively robust, even when the variables do not have a multivariate normal distribution, and maximizes the determinant of the partial correlation matrix among the variables with the common factors partialled out, regardless of the form of the distribution. In this case, the estimate of the unrotated factor pattern matrix is given by $\hat{\mathbf{A}} = \hat{\mathbf{\Theta}} \mathbf{A}_m [\gamma_i - 1]_m^{1/2}$, where $\hat{\mathbf{\Theta}}$ is the square root of the iteratively estimated diagonal unique variance matrix, $\mathbf{A}_m$ is the $p \times m$ matrix whose columns are the first m eigenvectors of $\hat{\mathbf{\Theta}}^{-1} \mathbf{R} \hat{\mathbf{\Theta}}^{-1}$, and $[\gamma_i - 1]_m$ is a diagonal matrix of order m whose diagonal elements are formed by subtracting 1 from each of the first m largest eigenvalues γ_i of $\hat{\mathbf{\Theta}}^{-1} \mathbf{R} \hat{\mathbf{\Theta}}^{-1}$. In this solution, the common factors are mutually uncorrelated. But the unrotated factor pattern matrix is only an intermediate solution. It defines a common factor space that contains the maximum common variance for any m common factors for these variables. It is not used for factor interpretation.

Step 5: Rotation of Factors

Mathematically, the factor pattern matrix is not unique. Given the model equation $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}$, we can find alternative factor pattern matrices and common factors by linear transformations of them. Let $\mathbf{T}$ be an $m \times m$ transformation matrix of full rank and $\mathbf{T}^{-1}$ its matrix inverse. Then $\mathbf{Y} = \mathbf{A}\mathbf{T}\mathbf{T}^{-1}\mathbf{X} + \mathbf{E} = \mathbf{A}^*\mathbf{X}^* + \mathbf{E}$, with $\mathbf{A}^* = \mathbf{A}\mathbf{T}$ and $\mathbf{X}^* = \mathbf{T}^{-1}\mathbf{X}$. This also has the form of a factor analysis equation. So, which factor pattern matrix should be our solution? L. L. Thurstone (1947) solved this problem with the idea of simple structure. He argued that if most of the observed variables were not functions of all of the common factors of their domain, then the vectors representing them in common factor space would fall in “coordinate hyperplanes”

or subspaces of the full common factor space. Each hyperplane or subspace would be spanned by, at most, $m - 1$ of the common factors, and all variables within those subspaces would be functions of just the common factors spanning the hyperplane. The common factors of the full common factor space then would be found at the intersections of these “coordinate hyperplanes,” because they would be among the basis vectors of each of $m - 1$ hyperplanes. If the coordinate hyperplanes could be identified by discovering subsets of the variables occupying these subspaces of lower dimension, then the common factors would be identified at their intersections. Furthermore, different selections of variables from the domain would still identify the same common factors at the intersections of the hyperplanes. Hence, the solution for the factors is not unique to the selection of variables from the domain and is an objective, invariant result over almost all selections of observed variables from the domain. To discover these hyperplanes, Thurstone proposed using hypothetical vectors called *reference axes* inserted into the common factor space. A subspace of the common factor space would be a subspace of vectors orthogonal to a reference axis and would be like a parasol whose ribs represent vectors in the subspace, with the reference axis its handle. Thus, one would move each reference axis around in the common factor space seeking different sets of variables to fall in the “parasol” orthogonal to the reference axis. These subsets of variables would define the coordinate hyperplanes. Originally, Thurstone used graphical, two-dimensional plots for each pair of factors and moved his reference axes manually in these two-dimensional spaces to discover sets of variables that would line up orthogonal to the reference axes. This was slow work. Later, with the advent of computers, analytic criteria for simple structure were formulated and the process of finding simple structure solutions automated. It is important, however, to realize that simple structure does not imply uncorrelated or orthogonal factors. The common factors may be correlated. Thus, one should avoid rotational procedures such as VARIMAX, which forces the common factors to be mutually orthogonal, if one wants a genuine simple structure solution. Good algorithms for simple structure that are generally available are direct Oblimin and Promax. Other algorithms claim a modest superiority over these, but they are not generally available in commercial factor analysis programs.

In our example, we rotated our four-factor solution to simple structure using direct Oblimin rotation. The

program produced a factor pattern matrix, a factor structure matrix, and a matrix of correlations among the factors. The factor pattern matrix contains weights like regression weights for deriving the observed variables from the common factors. The matrix is the most useful in interpretation of the factors, because these weights will be invariant under restriction of range across selected subpopulations of subjects. Furthermore, the pattern weights clearly show which variables are and are not functions of which common factors. The factor structure matrix contains the correlations between the observed variables and the common factors. It is itself the matrix product $\Lambda\Phi$ of the factor pattern and factor correlation matrices. So, once these two are obtained, the factor structure matrix is redundant. Furthermore, the factor structure matrix and the matrix of correlations among factors are not invariant under restriction of range or selection of subjects.

Step 6: Factor Interpretation

The method of factor interpretation is eliminative induction. One looks down each column of the factor pattern matrix for those variables having large “loadings” on the factor. One interprets the factor as that hypothetical variable that is common to those variables with large loadings but absent in variables with near-zero loadings. In Table 3, we find something like positive orientation to others or “kindness” to be what is common to those variables having high loadings on the first factor. Ability seems to be the common element for the second factor, whereas extraversion is the common

Table 3 Factor Pattern Loadings

	1	2	3	4
FRIEND	.78317	.04561	.06936	.02296
SYMPATH	.84525	-.05340	-.01319	.12206
KIND	1.03068	.02353	-.09171	-.04496
AFFECTN	.79470	-.03680	.14617	.03676
INTELL	-.07708	.88650	-.01461	.03989
CAPABL	.02454	.92586	.01284	-.02959
CMPTNT	-.07117	.90317	.00333	.10189
SMART	.10794	.91176	.00614	-.09064
TALKTV	.01144	.03235	.84101	-.09020
OUTGO	-.04548	-.00038	.92307	.08421
GREG	.01553	-.01468	.88010	.01090
EXTRAV	-.02696	-.01428	.93815	-.03029
HELPFL	.32864	.11450	.12744	.44996
COOPER	.23477	.10052	.11966	.56880
SOCIAB	.07860	.00542	.82472	.11774

Table 4 Correlation Between Common Factors

	1	2	3	4
FACTOR 1	1.00000			
FACTOR 2	.21857	1.00000		
FACTOR 3	.52371	.28226	1.00000	
FACTOR 4	.72944	.33731	.52443	1.00000

element for the third. The fourth factor indeed is something common to just helpful and cooperative, and these variables also have modest loadings as expected on kindness and ability. The correlations among factors in Table 4 suggest that kindness and extraversion share something in common and less with ability. In some cases, factor analysts will factor analyze the correlations among the factors to obtain second-order common and unique factors.

—Stanley A. Mulaik

REFERENCES

- Carlson, M., & Mulaik, S. A. (1993). Trait ratings from descriptions of behavior as mediated by components of meaning. *Multivariate Behavioral Research*, 28, 111–159.
- Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika*, 18, 277–296.
- Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, 19, 149–161.
- Guttman, L. (1956). “Best possible” systematic estimates of communalities. *Psychometrika*, 21, 273–285.
- Harris, C. W. (1962). Some Rao-Guttman relationships. *Psychometrika*, 27, 247–263.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441, 498–520.
- Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.
- Jones, M. B. (1959). *Simplex theory* (U.S. Naval School of Aviation Medicine Monograph Series No. 3). Pensacola, FL: U.S. Naval School of Aviation Medicine.
- Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32, 443–482.
- Jöreskog, K. G. (1979). Statistical models and methods for analysis of longitudinal data. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt.
- Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.
- Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research*, 22, 267–305.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 6(2), 559–572.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15, 366–374.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

FACTORIAL DESIGN

A factorial design involves the simultaneous study of two or more variables or factors on some other variable or variables. The factors may be manipulated or measured. Each factor consists of two or more levels or categories. The levels may differ qualitatively or quantitatively. For example, the qualitative factor of marital status may consist of the five levels of never married, married, separated, divorced, and widowed. Quantitative factors may comprise particular values (such as increasing dosages of a drug) or a range of values (such as the four age groups of 20–29, 30–39, 40–49 and 50–59).

A factorial design may be further described in terms of the number of factors, the number of factors and levels within each factor, whether cases have been randomized to factors or treatments, and whether cases are measured on more than one occasion. Factors may be referred to as “ways,” so that a two-factor design may be called a two-way design, a three-factor design a three-way design, and so on. Alternatively, the design may be described in terms of the number of levels within each factor. A design that consists of three factors, two having two levels and the third having four levels, may be designated a $2 \times 2 \times 4$ (“a two by two by four”) factorial design. A completely randomized factorial design is where each case has been RANDOMLY ASSIGNED to one and only one combination of factor levels or cells. A mixed factorial design comprises at least one factor where cases are measured on more than one occasion and one factor where cases are measured on only one occasion.

Ronald Fisher (1935) suggested that the factorial design had three advantages over the single-factor design. The first advantage is that it is more efficient or economical in that it requires fewer cases or observations for the same degree of precision or power. Compared with a single-factor design, a two-factor or two-way factorial design requires half as many cases, a three-way design a third as many cases, a four-way design a quarter as many cases, and so on (Snedecor, 1937). The reason for this is that the values

of one factor are averaged across the values of the other factors. The second advantage is that it is more comprehensive in that it allows the interaction between two or more factors to be examined. The third advantage is that it enables greater GENERALIZABILITY of the results in that a factor has been investigated over a wider range of conditions. A fourth advantage is that it may provide a more sensitive or powerful test of a factor if that factor does not interact substantially with one or more factors in that the other factors may account for some of the unexplained or error variance (Stevens, 2001).

—Duncan Cramer

See also EXPERIMENT

REFERENCES

- Fisher, R. A. (1935). *The design of experiments* (1st ed.). Edinburgh, UK: Oliver and Boyd.
- Snedecor, G. W. (1937). *Statistical methods* (1st ed.). Ames: Iowa State University Press.
- Stevens, J. (2001). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum.

FACTORIAL SURVEY METHOD (ROSSI'S METHOD)

Humans form ideas about the way the world works. They also form ideas about the way the world ought to work. These positive and normative ideas can be represented by equations, termed, respectively, *positive-belief equations* and *normative-judgment equations*. Rossi's factorial survey method makes it possible to estimate these equations-inside-the-head.

The positive-belief and normative-judgment equations are linked to two further equations—an equation describing the determinants of components of the beliefs or judgments, called a *determinants equation*, and an equation describing consequences of the belief/judgment components, called a *consequences equation*.

For example, individuals form ideas about earnings determination and, concomitantly, about the earnings they regard as just for themselves and others. Both the ideas about actual earnings determination and about just earnings determination are themselves the product of personal and social factors, such as information about the occupational structure and

religion. Moreover, the ideas about actual earnings help shape many decisions and behaviors—what to study, where to work, and so on—and the ideas about just earnings help shape many other decisions and behaviors—such as voting, joining a strike, making contributions to lobbying organizations, and designing a pay schedule.

Thus, the positive-belief equation and the normative-judgment equation each may join with a determinants equation to form a multilevel system of equations. Similarly, the positive-belief equation and the normative-judgment equation each may join with a consequences equation to form another (possibly multilevel) system of equations.

The factorial survey method pioneered by Peter H. Rossi (1951, 1979) and developed with associates (Jasso & Rossi, 1977; Rossi & Anderson, 1982; Rossi & Berk, 1985; Rossi, Sampson, Bose, Jasso, & Passel, 1974) provides an integrated framework for estimating the beliefs/judgments equations-inside-the-head, testing for interrespondent homogeneity, and estimating the determinants and consequences equations. All elements of the research protocol are designed with the objective of obtaining estimates with the best possible properties of the beliefs/judgments equations.

DATA COLLECTION IN THE FACTORIAL SURVEY METHOD

Each respondent is asked to rate the level of a specified outcome variable (e.g., healthiness, wage attainment, just prison sentence) corresponding to a fictitious unit (a person, say, or a family) that is described in terms of potentially relevant characteristics such as age, gender, study or eating habits, access to medical care or housing, and the like. The descriptions are termed “vignettes” (VIGNETTE TECHNIQUE). One of Rossi's key insights was that fidelity to a very rich and complex reality can be achieved by generating the POPULATION of all logically possible combinations of all levels of potentially relevant characteristics and then drawing random samples to present to respondents. Accordingly, the vignettes are described in terms of many characteristics, each characteristic is represented by many possible realizations, and the characteristics are fully crossed. Three additional important features of the Rossi design are the following: (a) In the population of vignettes, the correlations between vignette characteristics are all zero

or close to zero, thus reducing or eliminating problems associated with MULTICOLLINEARITY; (b) the vignettes presented to a respondent are under the control of the investigator (i.e., they are “fixed”), so that endogeneity problems in the estimation of positive-beliefs and normative-judgments equations arise only if respondents do not rate all of the vignettes presented to them; and (c) a large set of vignettes is presented to each respondent (typically 40 to 60), improving the precision of the obtained estimates. The rating task reflects the outcome variable, which may be a cardinal quantity, a probability, a set of unordered categories, or a set of ordered categories.

DATA ANALYSIS IN THE FACTORIAL SURVEY METHOD

The analysis protocol begins with inspection of the pattern of ratings, which in some substantive contexts may be quite informative (e.g., the proportion of workers judged underpaid and overpaid), and continues with estimation of the beliefs/judgments equation. Three main approaches are the following: (a) The classical analysis of covariance approach, in which the respondent-specific equations may have different error variances; (b) the seemingly unrelated regressions approach, in which the errors from the respondent-specific equations are correlated; and (c) the random coefficients approach, in which the respondents constitute a random sample, and the parameters of the respondent-specific equations are viewed as coming from a distribution. Under all approaches, an important step involves testing for interrespondent homogeneity.

Depending on the substantive context and on characteristics of the data, the next step is to estimate the determinants equation and the consequences equation. Again depending on the context, the determinants equation may be estimated jointly with the beliefs/judgments equation.

Rossi's method was, for many years, somewhat difficult to implement, given the computational resources required to generate the vignette population and to estimate respondent-specific equations. Recent advances in desktop computational power, however, render straightforward vignette generation, the drawing of random samples, and data analysis, thus setting the stage for a new generation of studies using Rossi's method.

FURTHER READING

Rossi's articles cited above, together with recent factorial survey studies, provide comprehensive exposition of procedures of data collection and data analysis.

—Guillermina Jasso

REFERENCES

- Jasso, G., & Rossi, P. H. (1977). Distributive justice and earned income. *American Sociological Review*, 42, 639–651.
- Rossi, P. H. (1951). *The application of latent structure analysis to the study of social stratification*. Unpublished doctoral dissertation, Columbia University.
- Rossi, P. H. (1979). Vignette analysis: Uncovering the normative structure of complex judgments. In R. K. Merton, J. S. Coleman, & P. H. Rossi (Eds.), *Qualitative and quantitative social research: Papers in honor of Paul F. Lazarsfeld* (pp. 176–186). New York: Free Press.
- Rossi, P. H., & Anderson, A. B. (1982). The factorial survey approach: An introduction. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments: The factorial survey approach* (pp. 15–67). Beverly Hills, CA: Sage.
- Rossi, P. H., & Berk, R. A. (1985). Varieties of normative consensus. *American Sociological Review*, 50, 333–347.
- Rossi, P. H., Sampson, W. A., Bose, C. E., Jasso, G., & Passel, J. (1974). Measuring household social standing. *Social Science Research*, 3, 169–190.

FALLACY OF COMPOSITION

Fallacy of composition is the failure to accept that what is true for the parts of a whole may not be true for the whole. A special case of this is referred to as the Tragedy of the Commons, where land held in common belongs to everyone collectively rather than to any one particular person. Thus, no one wants to take care of the property, and the land suffers. An everyday example is to reason that if one can see the ballet better by standing on one's seat, then it follows that if everybody stood on his or her seats, we would all have a better view.

—Tim Futing Liao

FALLACY OF OBJECTIVISM

This is the problem of falsely assuming that an entity exists independently of our consciousness. In PHENOMENOLOGY, the analyst seeks to bracket out such

assumptions. In the social sciences, it may take the form of falsely assuming that our conceptual categories exist, as opposed to being ways of understanding social reality.

—Alan Bryman

FALLIBILISM

Fallibilism is the deliberate and consistent attempt by researchers to disprove the fallibilities (both methodological and substantive) inherent in the research they are conducting. After the most rigorous scrutiny possible, if some methodological approaches or conclusions remain relatively intact, then researchers feel confident in the RELIABILITY and VALIDITY of these aspects of the research. Fallibilism thus views research as an engagement in continuous critical scrutiny.

As a research approach, fallibilism draws on three intellectual traditions. From the philosophy of science, it builds on the principle of FALSIFIABILITY. As articulated by Karl Popper (1965), falsifiability requires researchers to expose forms of inquiry, and specific HYPOTHESES, to the most rigorous attempt to disprove their central contentions. From this point of view, research is a Darwin-like search for the fittest knowledge. The emphasis is on disconfirmation, on probing for error, omissions, and contradictions.

From PRAGMATISM, fallibilism borrows the idea that inquiry is not a search for final universal truth, but a contingent, future-oriented activity in which researchers accept that the best knowledge available at any one moment is always open to further reformulation. Pragmatism has an EXPERIMENTAL, antifoundational tenor that advocates using an eclectic range of methods and approaches to promote democracy. It encourages the constant critical analysis of assumptions and is skeptical of reified, standardized models of practice or inquiry. The element of fallibilistic self-criticality endemic to pragmatism means avoiding a slavish adherence to a particular methodology.

From CRITICAL THEORY, fallibilism takes that tradition's skepticism regarding dominant ideologies. Critical theory is interested in laying bare the way that ideology—the system of ideas that seems normal, natural, and obvious and that explains the workings of society to people—helps maintain the power of an unrepresentative minority. Its main intellectual

project—ideology critique—seeks to expose the way that people learn to embrace ideas and practices that end up harming them, without their realizing that this is happening. Like fallibilism, critical theory assumes that what people view as universal norms, uncontested beliefs, and taken-for-granted explanations are, in fact, humanly constructed and infinitely malleable. Although fallibilism does not share critical theory's concern to expose the way in which dominant ideology legitimizes an unjust social order, it does share its skepticism regarding conventional truth.

A fallibilistic orientation always seeks to expose what is unwarranted, unexamined, and contradictory regarding a research method or set of findings. A good example of such a research effort is Gary Cale's (2001) analysis of how the attempt to encourage critical, oppositional thinking among community college students actually served to increase their resistance to this process.

—Stephen D. Brookfield

REFERENCES

- Bronner, S. E., & Kellner, D. M. (1989). *Critical theory and society: A reader*. New York: Routledge.
- Cale, G. (2001). *When resistance becomes reproduction: A critical action research study*. Proceedings of the 42nd Adult Education Research Conference. East Lansing: Michigan State University.
- Cherryholmes, C. H. (1999). *Reading pragmatism*. New York: Teachers College Press.
- Popper, K. E. (1965). *The logic of scientific discovery*. New York: Harper & Row.

FALSIFICATIONISM

Falsificationism is a philosophy of science, also known as critical rationalism, that uses the logic of deduction to provide the foundation for the HYPOTHETICO-DEDUCTIVE METHOD. It was developed in the 1930s by Karl Popper (1959) to deal with the deficiencies of POSITIVISM. The positivist position is rejected in favor of a different logic of explanation based on a critical method of trial and error in which theories are tested against "reality." Falsificationism shares some aspects of positivism's ONTOLOGY but rejects its EPISTEMOLOGY. It adopts the position that the natural and social sciences differ in their content but not in the logical form of their methods.

Although Popper was not a member of the Vienna circle, the group of scientists and philosophers that was responsible for developing logical positivism, he had a close intellectual contact with it. He shared with this tradition the view that scientific knowledge, imperfect though it may be, is the most certain and reliable knowledge available to human beings. However, he was critical of positivism—particularly logical positivism—and was at pains to distance himself from the circle. He rejected the idea that observations provide the foundation for scientific theories, and he recognized the important historical role played by metaphysical ideas in the formation of scientific theories.

Popper argued that observation is used in the service of deductive reasoning; theories are invented to account for observations, not derived from them, as in INDUCTION. Rather than scientists waiting for nature to reveal its regularities, they must impose regularities on the world and, by a process of trial and error, use observation to try to reject false theories (deduction). Theories that survive this critical process are provisionally accepted but never proven to be true. All knowledge is tentative and subject to ongoing critical evaluation. This critical attitude makes use of both verbal argument and observation; observation is used in the interest of argument.

The question of whether theories or observations come first was not a problem for Popper. He accepted that a hypothesis is preceded by observations, particularly those that it seeks to explain. However, these observations presuppose a frame of reference, that is, one or more theories.

In addressing the appropriate logic for the social sciences in his later work, Popper (1961) summarized what he called his main thesis. The logic of the social sciences, like that of the natural sciences, consists in trying out tentative solutions to certain problems; solutions are proposed and criticized. A solution that is not open to criticism must be excluded as unscientific. The ultimate criticism is to attempt to refute the theory. If the theory is refuted, another must be tried. However, if the theory withstands the testing, we can accept it temporarily, although it should be subjected to further discussion and criticism. Hence, the method of science is one of tentative attempts to solve problems by making conjectures that are controlled by severe criticism. The so-called objectivity of science lies in the objectivity of the critical method.

Popper argued that although science is a search for truths about the world or universe, we can never establish whether these theories are true. All that can be done is to eliminate false theories by a process of conjecture and refutation. Some theories will be rejected and some tentatively accepted (corroborated). This process allows us to get as near the truth as possible, but we never know when we have produced a true theory. Some scientist in the future may test the theory in some other circumstances and find it to be false. Therefore, theories must always be regarded as tentative—they may be refined or refuted in the future. All that we can do is get rid of theories that do not match reality. Popper believed that there is no more rational procedure than the method of conjecture and refutation. It is a process of boldly proposing theories, of trying our best to show that these are wrong, and of accepting them tentatively if our critical efforts are unsuccessful.

Criticisms made of the logic of deduction also apply to falsificationism.

—Norman Blaikie

REFERENCES

- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Chalmers, A. F. (1982). *What is this thing called science?* St. Lucia: University of Queensland Press.
- Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.
- Popper, K. R. (1961). *The poverty of historicism*. London: Routledge & Kegan Paul.

FEMINIST EPISTEMOLOGY

Feminist epistemology consists of principles of EPISTEMOLOGY that have been informed by feminism.

—Alan Bryman

See also FEMINIST RESEARCH, STANDPOINT EPISTEMOLOGY

FEMINIST ETHNOGRAPHY

ETHNOGRAPHY that has been informed by the principles of FEMINIST RESEARCH. As such, feminist ethnography documents women's lives and

activities, conducts research and analysis from women's perspectives, and appreciates women's positions and perspectives contextually. There has been some debate about the prospects for feminist ethnography on the grounds that it is difficult to conduct truly according to feminist principles, because it is difficult to conduct in a totally nonexploitative way. Against such a viewpoint, feminist researchers argue that injecting reciprocity into the interactions and transactions between researchers and participants can help to offset an exploitative relationship between researchers and the women they study.

—Alan Bryman

FEMINIST RESEARCH

Feminist research developed as a response to two perceived related failings in western social sciences. The first was the relative invisibility of women and a lack of concern with the gender-specific issues that influenced their lives. The second concerned the practices of social research and the processes through which knowledge was constructed. It was argued that the social world had been studied from the perspective of male interests and concerns, and in ignorance of the different picture that emerged when focusing on women's lives and ways of seeing. Knowledge, which was presented as neutral, objective, and value-free, was, instead, partial and gendered. The purpose of feminist research was to bring women's experiences more fully into view. This necessitated challenging conventional research practices, as well as radically reviewing many of the taken-for-granted assumptions about the nature of social science. The overall aim was to understand better the nature of gender inequalities.

It is not the focus on women, gender, or gendered lives per se, however, that makes a research project feminist. At first, feminist research was defined as such if it was seen to be about, by, and for women. More recently, however, it has been suggested that feminist approaches to research can be identified through their framing by theories of gender and power, their normative frameworks and notions of justice (however conceived), their focus on transformation and social change, and ideas about ethics and accountability (Ramazanoglu, 2002). Views about these issues are not held uniformly, and identification of when research

might be feminist is never an open-and-shut case. There is also overlap between the defining features of feminist work and other approaches to social investigation.

Early on in their discussions about doing social research, feminists distinguished between method, methodology, and EPISTEMOLOGY (Harding, 1987). Method was used to refer to the techniques and tools used in conducting research. Methodology was concerned with how those methods were actually operationalized and put into practice. Epistemology focused on the philosophical arguments for deciding what kinds of knowledge are possible and how to ensure that they are adequate and legitimate. Feminists have continually refined their ideas about these issues, and there has been extensive debate in relation to them.

METHODS OF RESEARCH

During the early 1980s, much of the debate about feminist research was concerned with what constituted the appropriate methods to be used. Concerns were expressed about the greater legitimacy that seemed to be afforded to SURVEY research and to approaches that emphasized the importance of scientific MEASUREMENT and OBJECTIVITY at that time. It was argued that such research fractured people's lives, resulting in the production and measurement of atomistic facts, the significance of which had been decided in advance of the research itself. This was seen as problematic when focusing on women and, particularly, for research into those areas of their lives that had remained relatively hidden. With issues such as violence, sexuality, and motherhood, for instance, there was little existing knowledge from which survey questions might be written.

As a result, many feminists argued that qualitative methods, founded on an INTERPRETIVIST philosophical position and using sensitive, flexible, and open-ended approaches to data generation, were more appropriate than quantitative ones. Methods such as IN-DEPTH INTERVIEWS, LIFE HISTORY INTERVIEW, and ETHNOGRAPHY focused on the meanings and interpretations of those being researched, thereby enabling the researcher to see the social world through participants' eyes. It was also felt that qualitative approaches were more in keeping with feminist principles, which promote equality and are against objectification and subordination. These arguments became something of an orthodoxy among feminist researchers, although even then, not everyone agreed with this position. For instance,

awareness of the prevalence of sexual violence or the poverty of lone mothers is dependent, in part, on the availability of data that can indicate the extent of the problem.

In recent years, the old orthodoxy about QUALITATIVE RESEARCH has been breaking down. Some commentators have claimed that such methods may be more exploitative than quantitative ones because they rely on assumptions about trust and RAPPORT, which may mask relations of power and inequality. Feminists have been adopting a more pragmatic approach toward mixing methods and even adopting quantitative ones in a bid to develop a more emancipatory social science that can inform policy development. Indeed, some have suggested that there may be circumstances (e.g., investigating sexual abuse) where respondents may find completing a questionnaire to be less stressful than being interviewed. However, limitations of funding and resourcing often mean that relatively small-scale qualitative projects remain the main way in which many feminists are able to contribute to the research agenda.

METHODOLOGY

The debate about methods of research has been related to a second issue concerning feminist methodology or research practice. It has been suggested that what distinguishes feminist research from other forms has less to do with methods per se and more to do with the kinds of questions that are asked, the relationship of the researcher to the research and its conduct, and the intentions and purpose behind the work. There has been particular concern about how to ensure that research takes place in nonexploitative and ethically responsible ways. A central issue here has been the possible existence of structural inequalities between researchers and their participants and the extent to which these might be minimized. It has been argued that feminists who replicate, during their research, the kind of power relations of which they are critical elsewhere are hypocritical, and such action undermines the knowledge produced.

A range of strategies for dealing with such issues has been debated, with some filtering into the more mainstream practices of nonfeminist research. These have involved, for instance, developing good rapport and communication with participants in research by telling them as much as possible about the content and aims of the research; responding to participants' questions, problems, and concerns as honestly and

sensitively as possible; and making available the names of organizations or publications that can provide any necessary information or social support. Some feminists advocate obtaining feedback from the researched on the interpretation and analysis of research data in order to ensure that women's voices are heard without distorting or exploiting them. Emphasis has been placed on including participants in the design and implementation of a research project, as well as ensuring that findings are disseminated to those who have been involved.

However, although some feminist researchers seem to assume that it is possible to remove most aspects of power relations from the research process, many are less sanguine. It may be all too easy for an educated researcher to persuade participants to talk conversationally, in relaxed circumstances, about their lives. But with the benefit of hindsight, they may feel it would have been better to remain silent. It is also easy for researchers to believe that they have established relations of harmony and equality, when the perceptions of those whom they are studying is otherwise. Some feminists have also suggested that the power of researchers can be overstated, with the researched being ascribed a passive, victimized status, where they have no apparent space for resistance.

One way in which the issue of power in research has been addressed is through the concept of REFLEXIVITY, which has become something of a central principle in feminist methodological concerns. Reflexivity has two distinct aspects. The first involves making explicit how power might be exercised in and affects the research process despite strategic efforts to minimize its impact. It focuses on the assumptions and ethical judgments that frame the research and how the research agenda and process have been constructed. It thereby demonstrates how researchers can account for the knowledge they produce (Ramazanoglu, 2002). The idea is to turn the critical methods of social researching upon the practice of social researching itself.

The second form of reflexivity explores the social situatedness of the researcher. This refers to how the personal biography and position of the researcher relate to the project in which she is engaged. The argument is that researchers cannot be separate from their work, which has no existence apart from their involvement in it. Therefore, it is important for a critical reflection on this relationship to be included as part of any discussion, analysis, or interpretation of research findings.

In fact, feminists have claimed that this two-fold reflexive commentary on research is essential to making knowledge accountable. They are critical of what is referred to as "weak" objectivity, where attempts are made to purge knowledge, along with the research on which it is based, of all values (Harding, 1991). This form of objectivity is weak because, by pretending that such values do not exist in research, they are included in hidden and unexplicated ways. In contrast, what is known as "strong" objectivity necessarily involves critical scrutiny of all evidence marshalled as part of the research process (Harding, 1991). Strong objectivity, with its emphasis on reflexivity, includes the systematic examination of background beliefs and cultural agendas. By making this part of the research process itself, it is argued that the knowledge produced becomes more rigorous, robust, and defensible.

There is debate, however, as to how much reflexivity and strong objectivity might actually be achieved. Being reflexive in the ways suggested can be a long and drawn-out business, with difficulty in establishing guidelines as to how far or for how long the process should continue. There are dangers that accounts of the research procedure and of researchers' involvement come to dominate the analysis and findings of the research. This could result in the complexities of participants' lives being reduced to the researcher's own autobiographical history.

One issue that has significantly affected feminists' views on methodology relates to the concept of difference. Early assumptions that there is an essential womanhood and that women's oppression is universal, homogeneous, and shared have been challenged. An emphasis on difference draws attention to diversities among women, particularly those deriving from ethnicity, sexuality, disability, class, and age. This means that just being a woman researching women cannot be sufficient in the effort to establish rapport and minimize hierarchy. These other axes of experience will also impinge on the research process, and this is why it has been suggested that researcher and researched might be matched in terms of shared characteristics. However, it is not always easy to establish which differences are likely to have an influence in a particular study and what the precise nature of this impact might be. Overall, then, shared gender characteristics do not themselves promote shared understandings, and it cannot be assumed that trust can be established easily on the basis of gender alone.

EPISTEMOLOGY

The feminist concern with epistemology has centered on who knows what, about whom, and how this knowledge is legitimated. Early writers were critical of those who simply conducted their studies of women within already existing frameworks. In order to make the meaning of women's lives more visible, it was necessary to analyze it from their point of view. This has come to be known as STANDPOINT EPISTEMOLOGY (Hartsock, 1998). There are different ways of thinking about taking a feminist standpoint, but the key is that women's particular gendered experiences produce distinctive and privileged understandings. This access to different knowledge makes it possible to reveal the existence of forms of human relationships that may not be visible from a man's position. Therefore, standpoint epistemology offers the possibility of new and more reliable insights into women's lives because it is grounded in women's experiences, including emotions and embodiment.

Standpoint epistemology is also based upon a REALIST approach to knowledge; the idea that systematic or structural factors influence LIVED EXPERIENCE, and that these structures are knowable. The whole point of feminist research has been to produce continually better understandings of women's lives. This view, however, has been challenged by developments in postmodern thinking, and the standpoint perspective is often contrasted with postmodern accounts.

As with all forms of POSTMODERNISM, feminist thinking has different forms and emphases. However, it shares most of the characteristics of other (non-feminist) variants. There is the criticism of generalizations, universal categories, the existence of a stable and coherent self, the transparency of language, and the ability of scientific rationality to produce "truth" (Nicholson, 1990). The focus, instead, is on discourse, textuality, fragmentation, multiple subjectivities, and flux. Whereas the standpoint position acknowledges that accounts are selective constructions, while also agreeing that they are able to represent independent phenomena with some degree of rigor and reliability, feminist postmodernism rejects this. Instead, some have claimed that feminist postmodernism sweeps away the foundations of feminist methodology (Ramazanoglu, 2002). It does this by arguing that because social phenomena are only to be apprehended through the use of discourse, and because the very practice of discourse serves to invoke forms of sociality, it is therefore impossible to understand the

social world without also simultaneously constructing it. The latter then becomes the task of what it means to "know." The emphasis switches from trying to investigate the nature of social relations from accounts of women's experiences to the discourse and texts through which things come to be known. Indeed, the very idea of experience itself, as something that people have, and that can be taken at face value unquestioningly, becomes problematic. All of this causes major difficulties for social researchers who wish to explore critically the nature of the social world; from postmodern perspectives, this is impossible. While not denying their own material embodiment or that they live in a material world, feminist postmodernists claim that particular links between experience and material realities cannot be specified (Ramazanoglu, 2002). Rather, feminist research is transformed into the continuing deconstruction of identities, subjectivities, and selves, which is then subjected to further deconstruction.

Postmodern thinking has had a considerable impact on feminist research, with some regarding any attempts at any empirical work as misguided and defective. Others argue that it is possible to use some of the insights to be gained from a postmodern approach, particularly when researching culture, signification, and modes of representation. The problem for the feminist social researcher is how to strike a balance between an empirical focus on embodied and material differences, power, and inequality, and critical reflections on how knowledge is produced. The question as to whether it is possible to combine concerns about the relationship between subjectivity and the production of texts/interpretations with empirical methods of social investigation is still being debated. It produces differences among feminists, who respond with various and divergent answers.

—Mary Maynard

REFERENCES

- Harding, S. (Ed.). (1987). *Feminism and methodology*. Milton Keynes, UK: Open University Press.
- Harding, S. (1991). *Whose science? Whose knowledge?* Milton Keynes, UK: Open University Press.
- Hartsock, N. C. M. (1998). *The feminist standpoint revisited and other essays*. Boulder, CO: Westview.
- Nicholson, L. (Ed.). (1990). *Feminism/postmodernism*. London: Routledge.
- Ramazanoglu, C., with Holland, J. (2002). *Feminist methodology*. London: Sage.

FICTION AND RESEARCH

Until the mid-20th century, social research tended to reduce fiction to determining the social contexts of its literary production—a POSITIVIST legacy in evolutionism, structural-functionalism, and Marxism. An alternative, *Literaturwissenschaft* approach, analyzes fiction for knowledge about the society of its production and to which it refers, and informs analyses of fiction by cultural studies (Milner, 1996). It combines two distinct approaches: Extrinsic approaches analyze fiction as a critical reflection on the sociocultural conditions of its production, distribution, and reception; intrinsic approaches analyze it as a reflexively constitutive feature of its society, in which linguistic structures of TEXTS are analyzed as structural homologues of social relations. Intrinsic approaches seek to avoid reductions of fiction to illustrations of social realities by emphasizing its textual linguistic subtleties and complexities, but they raise research questions of representativeness of majority experience in complex modern societies and are applied more effectively to the fiction of both minority, selective cultural traditions and avant garde modernism. Extrinsic approaches provide more effective analyses of popular fiction because of their lack of sustained comparative concern with social experience in other cultures and periods (Loewenthal, 1989).

More recent sociolinguistic and neo-Marxian STRUCTURALIST research on relations between language and social structure as constitutive of texts (Filmer, 1998) has focused on the HERMENEUTIC issue of how the language of fiction indicates possible readings. This requires openness to a text's capacity to invent both itself and the reality to which it refers. The following approaches address this.

Semiotics and Structural Linguistics

SEMIOTICS and STRUCTURAL LINGUISTICS provide theories of signs required to make sense of the language of the text and its relations to languages of everyday life and other critical discourses (Eco, 1981).

Phenomenological Sociology

Phenomenological sociology analyzes relations between meanings of idiomatic expressions and connotations contained within the text's version of reality

and those of the external realities of readers' worlds of experience and discourse (Schutz, 1964).

Structures of Feeling

Structures of feeling analyzes the underlying symbolic and semantic organization of texts as expressions of emergent experiential senses of changing contexts, prefiguring possible new social structures (Prendergast, 1995).

Genetic Structuralism

Genetic structuralism applies structuralist Marxist analysis of fiction to articulate world visions of emergent social groups. This shared consciousness is analyzed for its potential in enabling the groups developing it to change the existing institutional social and political order (Goldmann, 1981).

All of these approaches engage with fiction as a reflexive relation between language, subjective imagination, and social structures to establish the adequacy of fiction as a resource on fundamental topics of social research: structures of relations between individuals and society, processes of social change, and meanings as well as causes of social action. In doing so, they seek to transcend the specificities of personal biographies, particular cultures, and historical periods.

—Paul Filmer

REFERENCES

- Eco, U. (1981). *The role of the reader: Explorations in the semiotics of texts*. London: Hutchinson.
- Filmer, P. (1998). Analysing literary texts. In C. Seale (Ed.), *Researching society and culture*. London: Sage.
- Goldmann, L. (1981). *Method in the sociology of literature* (W. Boelhower, Trans. & Ed.). Oxford, UK: Basil Blackwell.
- Loewenthal, L. (1989). Sociology of literature in retrospect. In P. Desan et al. (Eds.), *Literature and social practice*. Chicago and London: University of Chicago Press.
- Milner, A. (1996). *Literature, culture and society*. London: UCL Press.
- Prendergast, C. (1995). *Cultural materialism: On Raymond Williams*. Minneapolis: University of Minnesota Press.
- Schutz, A. (1964). Don Quixote and the problem of reality. In *Collected papers*, Vol. 2. The Hague: Martinus Nijhoff.
- Williams, R. (1977). *Marxism and literature*. Oxford, UK: Oxford University Press.

FIELD EXPERIMENTATION

Field experiments, as distinct from LABORATORY EXPERIMENTS, are randomized interventions that take place in naturalistic settings. The nature of these interventions varies widely. Examples include randomized school voucher programs, voter mobilization campaigns, police raids on crack houses, job training programs, income tax schedules, and health insurance plans. The unit of analysis in field experimentation also varies. Most field interventions target individuals, but many studies randomly assign treatments to groups or institutions.

In contrast to ethnographic or descriptive research, field experiments are principally designed to establish CAUSAL relationships. Well-executed field experiments combine the strengths of randomized designs with the EXTERNAL VALIDITY of field studies. Whereas a laboratory study might examine the effects of commercial advertising by exposing TREATMENT and CONTROL GROUPS to different advertising stimuli within an artificial setting and gauging their purchasing preferences within the context of a simulated marketplace, field experiments randomly manipulate the content and timing of actual advertisement campaigns and attempt to link these varied interventions to observed patterns of consumer demand. Both types of studies use randomization, but the latter has the advantage of linking cause and effect in terms that have direct, real-world applicability.

Field experimentation is also a valuable tool in program evaluation, particularly when used to gauge the effects of a new program or alternative versions of existing programs. Suppose, for example, that one sought to evaluate the effectiveness of school-based drug prevention programs. Some schools might receive drug awareness programs and others might not, the sequence in which schools receive drug awareness programs might be varied randomly, or the entire population of schools might receive alternative drug awareness programs. By comparing outcomes in treatment and control groups that were initially formed by random chance, the evaluator may make an unbiased assessment of the intervention's effects.

One attractive feature of experiments in natural settings is the transparency of the data analysis. In contrast to nonexperimental data analysis, where results often vary markedly depending on the model the researcher imposes on the data and where researchers often fit a

great many models in an effort to find the right one, experimental data analysis tends to be quite robust. Simple comparisons between control and treatment groups often suffice to give an unbiased account of the treatment effect, and additional analysis with COVARIATES merely estimates the causal parameters with greater precision. This is not to say that experimental research is free from data mining, but the transformation of raw data into statistical results involves less discretion and therefore fewer moral hazards.

In principle, field experimentation represents the strongest basis for sound causal inference available to social scientists, but in practice, this methodology faces important ethical and pragmatic constraints. Rarely do social scientists have the opportunity to manipulate the variables of most interest to them, such as culture, political systems, economic prosperity, and so on. Even in situations where random interventions are attempted, they are sometimes undone by the people charged with implementing them. The well-intentioned physician who smuggles the sickest patients into the treatment group may cause the experiment to produce misleading results.

The challenge of orchestrating field experiments and maintaining the integrity of the randomization means that the experiments tend to occur in a small number of sites that are chosen for reasons of logistics rather than representativeness. This constraint raises the issue of whether the study's conclusions apply only to the types of people who actually participate in an experiment. In general, REPLICATION is the appropriate response to concerns about drawing conclusions based on studies of particular times, places, and people.

A further complication arises when experimental subjects refuse to participate in the study or cannot be reached for treatment. Although noncompliance and attrition diminish the power of an EXPERIMENTAL DESIGN, they are remediable problems as long as the decision to participate is unrelated to the strength of the treatment effect. The statistical correction is to perform an instrumental variables regression in which the independent variable is whether a subject was actually treated and the INSTRUMENTAL VARIABLE is whether a subject was originally assigned to the TREATMENT group. The resulting effect estimate is termed the "effect of treatment on the treated." Whether the effects of the treatment do, in fact, vary may be studied empirically by experimentally manipulating participation rates (e.g., by varying inducements offered to subjects).

In light of these practical and ethical concerns, social scientists often turn to NATURAL EXPERIMENTS, where treatment and control groups are formed in ways that resemble random assignment. For example, if municipal leaders increase the size of the police force during election years in an effort to placate a crime-conscious citizenry, the electoral cycle provides a means for studying the effects of policing on crime rates. Similarly, if public schools admit students to kindergarten on the basis of their birthdates, one may study the effect of kindergarten enrollment on intellectual development by comparing children whose birthdays fall immediately before and after the cut-off date. Like field experimentation, these approaches seize upon opportunities to learn about the consequences of near-exogenous variation in naturalistic settings. Unlike experimentation, the exogeneity of the independent variables is not ensured by means of randomization procedures.

—Donald P. Green

REFERENCES

- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444–455.
- Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution.
- Green, D. P., & Gerber, A. S. (2002). Reclaiming the experimental tradition in political science. In I. Katznelson & H. V. Milner (Eds.), *Political science: State of the discipline* (3rd ed., pp. 805–832). New York: W. W. Norton.
- Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, 9, 85–110.
- Rosenzweig, M. R., & Wolpin, K. I. (2000). Natural “natural experiments” in economics. *Journal of Economic Literature*, 38, 827–874.

FIELD RELATIONS

The definitive characteristic of social science field research is the encounter between researcher(s) and researched. Encounters in the field may vary from those that are short-lived to those that are sufficiently long-lived to form the basis of ongoing relationships. In the latter case, the nature of relations

between the researcher(s) and the researched is central to the progress and success—or otherwise—of the research.

Perhaps the most obvious point is that all intensive field research requires the careful and reflexive cultivation of ongoing good relationships with a wide range of people. Two of the most important field relations are with *gatekeepers* and *key informants*. These are vital—indeed, probably unavoidable—if long-term field research is to be successful. Gatekeepers facilitate initial ACCESS to the field and to information, situations, and individuals within it, but their continued sponsorship remains important throughout the research period. Typical gatekeepers include politicians, archivists, senior managers within organizations, community leaders, and core members of informal networks. Key informants, of whom the archetypal example is William Foote Whyte’s “Doc” in *Street Corner Society* (Whyte, 1955), may also be gatekeepers, and often are, but the reverse is not necessarily true. Key informants and gatekeepers become disproportionately significant if the fieldworker lacks working fluency in the local language.

Among the problems raised by the importance of gatekeepers and key informants are how to identify the most useful gatekeepers; how to prevent oneself from becoming overidentified with gatekeepers and key informants; how to prevent gatekeepers from exercising proprietary control over the research; and how to avoid privileging the accounts and knowledge of a few, typically not disinterested, key informants over the points of view of other research subjects. None of these has a straightforward, textbook solution.

Relationships in the field are particularly important when the researcher has an already existing *membership role* within the social setting under investigation (Adler & Adler, 1987). There is much to gain in terms of privileged access and understanding from insider status, whether that be total or partial, and the realities of much social research today, particularly evaluative work, are that this is increasingly common. However, in such situations, it is possible to identify a number of issues that demand attention, including maintaining critical distance and objectivity, dealing with the demands made by other members in terms of their personal or corporate interests, balancing ethical responsibilities, and managing subsequent reentry into the situation as a nonresearcher.

Finally, perhaps the most pervasive and routine sense in which relations between fieldworker(s)

and research subjects are consequential for research outcomes reflects the crucial importance of *gender* in the everyday human world (Warren & Hackney, 2000). This is not surprising, given that everyday life, no matter what the setting or culture, is organized around gender. Who can go where, with whom, and who can speak to whom are gender issues that are critically important for the conduct and outcome of all field research, whether the topic of the research is gender-oriented or not.

—Richard Jenkins

REFERENCES

- Adler, P., & Adler, P. A. (1987). *Membership roles in field research* (Qualitative Research Methods Series vol. 6). Newbury Park, CA: Sage.
- Warren, C. A. B., & Hackney, J. K. (2000). *Gender issues in ethnography* (Qualitative Research Methods Series vol. 9, 2nd ed.). Thousand Oaks, CA: Sage.
- Whyte, W. F. (1955). *Street corner society* (2nd ed.). Chicago: University of Chicago Press.

FIELD RESEARCH

The expression “field research,” or “fieldwork,” derives in the first instance from natural science, where it denotes the collection of data in a naturalistic setting that is outside the controlled environment of the laboratory. In our context, despite its recent overidentification with ethnography and other qualitative approaches (e.g., Burgess, 1982; Whyte, 1984), most social research can probably lay claim to the name. Whether it be the fleeting doorstep exchanges of the large-scale social SURVEY or the long-term intimacies made possible by PARTICIPANT OBSERVATION, social research involving primary data typically involves at least some fieldwork—defined simply as encounters between researchers and subjects in uncontrolled environments—of one sort or another (with the most significant exceptions being ARCHIVAL RESEARCH, media studies, and FOCUS GROUP studies).

Why should social researchers depend so heavily on fieldwork? Or, to ask the question more directly, why do social researchers rarely, if ever, engage in experiments? There is no single answer. The ethics of the broad sociological tradition may provide one key constraint to experimentation on humans (although the fact that ethically sensitive disciplines such as medicine and

psychology continue to rely on experimental evidence suggests that ethical issues are unlikely to be determinate). Three other factors are, perhaps, more to the point.

The first is the difficulty in creating sufficiently authentic social conditions out of their actual context (with the exception of certain small-scale and definitely bounded situations). Simulation is, after all, simulation, and it is probably of little relevance for our understanding of the everyday human world. Second—and it is a related point—it is difficult to maintain properly controlled experimental conditions in social research contexts. Each of these problems means that for most purposes, there is no social research substitute for NATURALISM.

Third, and perhaps most important, there is the matter of predictability. Humans are reflexive beings. Those aspects of human behavior with which social research is concerned derive from a mixture of emotion, rational decision making, habit, and compulsion. For this reason, humans cannot be relied upon to do in the future what they have done consistently in the past. Because the goal of the experimental method is to accumulate sufficient accurate observations of an event and the controlled conditions under which it occurs to permit reliable *prediction* (i.e., the formulation of probabilistic generalizations), it clearly has limited application in social research.

This is not to say, however, that there are no approximate analogues of experimentation in field research. In particular, STRUCTURED INTERVIEWS have something in common with the experimental method, in that the object of the exercise is to discover what a sample of subjects will say in answer to the same questions asked in the same controlled way, or as close as is possible. Rigorous sampling, too, may be seen as another proxy for experimental control. Furthermore, projective questions of the “What if?” kind have an essential experimental quality to them (they certainly have something in common with what philosophers call “thought experiments”).

Nevertheless, with its necessarily uncontrolled environment, and in the intractability of the human behavior with which social research is concerned, field research is very different from the laboratory experiment. Epistemologically, field research also has some distinctive characteristics. The most obvious is its utter reliance on researchers, rather than instruments, for the perception and recording of data. The researcher *is* the research instrument, even when using the most

structured schedules or protocols. As a consequence, it can be difficult to say where data collection finishes and its interpretation begins.

This also means that in field research, the line between method and data can sometimes blur. For example, when Howard Newby studied English farm workers, the fact that the only way he could find and contact a sample of laborers was via their employers was not merely an interesting detail of method, it was a telling indicator of their social isolation (Newby, 1977). Similarly, that female researchers typically have better access to the lives of young men than their male counterparts have to young women says a great deal about the asymmetrical character of public and private gendered, everyday worlds.

These examples suggest a further point: that the nature, extent, and boundaries of the field are not necessarily self-evident. Unless one is talking about research settings that are clearly delimited territorially and/or institutionally—as in the small settlements and workplaces that have hosted some of the classic ethnographies—then defining the field of study is one of the most important, and often least straightforward, preliminary tasks of fieldwork. In addition to the issues of sampling and gender already mentioned, size and complexity are necessary considerations. Such matters affect which methods are appropriate and what the costs are of adopting one approach rather than another. For example, most modern settings are too extensive, too heterogeneous, and too multifaceted to be easily accessible “in the round” via participant observation (Jenkins, 1984). Thorough fieldwork commonly demands a degree of flexibility and openness with respect to method.

What one can find out in the field also depends on the politics of the research process. With respect to politics, the key issues involve questions of researcher identity, the role of gatekeepers, and (over)reliance on particular sources. Who the researcher is, via whom access to the field is negotiated, and with whom researchers develop key informant relationships are all factors that will influence what can be discovered, to whom one can talk, who will respond to requests for cooperation, and which local points of view are represented (see FIELD RELATIONS).

Given the above issues, why do social researchers continue to undertake field research? To return to the point of departure, with a few notable exceptions, the collection of primary social research data allows for no alternatives. Field research is a continued necessity

if social research is to maintain a critical engagement with the human world and its changes. Archives, the media, and focus groups can contribute only so much to that task, after which there is no substitute, whatever methods we adopt, for encountering humans in their everyday settings.

—Richard Jenkins

REFERENCES

- Burgess, R. G. (Ed.). (1982). *Field research: A sourcebook and field manual*. London: Allen & Unwin.
- Jenkins, R. (1984). Bringing it all back home: An anthropologist in Belfast. In C. Bell & H. Roberts (Eds.), *Social researching: Politics, problems, practice* (pp. 147–164). London: Routledge and Kegan Paul.
- Newby, H. (1977). In the field: Reflections on a study of Suffolk farm workers. In C. Bell & H. Newby (Eds.), *Doing sociological research* (pp. 108–129). London: Allen & Unwin.
- Whyte, W. F. (1984). *Learning from the field: A guide from experience*. Beverly Hills, CA: Sage.

FIELDNOTES

Fieldnotes constitute the ETHNOGRAPHIC record of ethnographic or participant observation research and are arguably the most essential part of the field research enterprise. Comprising the chronological log of experiences in the field, they include descriptions of people, events, the fabric of the setting, conversations with people, observed interactions, and sequences and duration of events, as well as the researcher's experiences connected to the investigation. It is important, too, that the observer note anything about the general state of him- or herself that might have bearing on the form or quality of the data. These notes may also contain interpretations of observed events and interactions and insights into the culture studied. In short, fieldnotes comprise the contents from which the analysis derives.

Fieldnotes should be written up as completely and comprehensively as soon as possible. Researchers generally agree that fieldnotes are more vivid and detailed the sooner they are recorded. However, in many situations, it is impossible, inconvenient, or inappropriate to record observations immediately. In such cases, it is usual to make temporary notes first, which essentially provide a summary that could be used to write up more detailed notes later. Such condensed notes,

which typically include phrases, single words, and unconnected sentences, can be recorded quickly and even inconspicuously, and serve as memory triggers to help reconstruct an event. Then, the researcher can expand on this condensed version by filling in the details and recalling things that were not recorded on the spot. To be sure, field notes are comprehensive only in the relative sense. Even permanent notes are altered, added to, corrected, and updated as the researcher becomes more familiar with the range of data at his or her disposal and with emerging hypotheses and interpretations.

Labor-intensive and requiring considerable self-discipline, some fieldnotes may, at first, seem obvious and trivial. However, researchers should not assume in advance what kinds of information fall within this domain. At least initially, the observer's mind-set should be that everything occurring in the field is potentially important and, therefore, worth recording. As the research becomes more focused, the researcher is more selective in what is written down.

Because researchers are selective in what is observed and recorded, and the specific words and terminology used, the language employed necessarily, even if inadvertently, imposes a structure on the social world under study. To mitigate against the obtrusiveness of preconceptions, cultural and otherwise, researchers should record verbatim what people say and avoid writing summaries of people's remarks or restating them in a more familiar language.

Fieldnotes should be reviewed frequently because this process allows them to be corrected and completed. Such a review also assists in the tasks of data categorization and analysis, which are best pursued simultaneously with data collection. Thus, for example, fieldnotes may contain analytic MEMOS that complement those of a more substantive and methodological nature and that form the core of the preliminary analysis. Such memos may include summaries that are written at the end of a day in the field in which the researcher indicates themes that have emerged and concepts that might be developed, together with preliminary thoughts about the analytic framework. Moreover, the review also provides a guide to areas where additional data are required. If their contribution to the emerging analysis is not considered systematically, and if accumulated for their own sake, fieldnotes become progressively less useful and, perhaps, even useless.

Generating and organizing fieldnotes is a personal activity, so researchers rely upon a format that suits

their taste and requirements. Generally, it includes some variation of the following: Each set of notes begins with the date, time, and place of observation, and when the notes were recorded; where appropriate, pseudonyms are used for names of people and places; and verbatim descriptions of people, events, and situations should be distinguished from those that are paraphrased. Because the objective is to record information as correctly and quickly as possible, there is little need to be overly concerned with grammar and spelling. The literature includes stylistic suggestions distinguishing between verbatim accounts and those based on reasonable and approximate recall.

Proficiency at writing up fieldnotes is gained over time and with practice. Experience, however, does not diminish the reality that writing fieldnotes is tedious, requiring patience and perseverance. However, the work is not characterized solely by drudgery: Writing fieldnotes is frequently accompanied by flashes of insight, excitement, and understandings about the social world under investigation.

Although most observers traditionally rely on their memories for recording data, it is not unusual for researchers to use recording devices. Despite the absence of absolute rules regarding such practices, a general consensus is that the researcher's obtrusiveness may detract from the quality of the data that are collected, recording devices should be used advisedly, and then only after some measure of familiarity with the setting and its key participants has been achieved.

Prior to the advent of computer programs, the task of categorizing fieldnotes was arduous. Computer programs for qualitative data can now perform many of the tedious, time-consuming, and mechanical operations associated with traditional analytic procedures as well as allow for the more rapid retrieval of field data. These operations can be performed with greater efficiency and accuracy, thus allowing the researcher to devote more time to the interpretive phase of data analysis. However, in and of itself, the technology adds little to nothing to the organization and analysis of fieldnotes without the researcher's conceptual input.

Because a good ethnography is only as good as the field notes upon which it is based, it is surprising that published accounts of field research methods pay relatively little attention to how the notes are written and organized. Their importance is underscored by the generally accepted view that field researchers may as

well not spend time in the research setting if they fail to record their fieldnotes in a timely manner.

—William Shaffir

REFERENCES

- Berg, B. L. (2001). *Qualitative research methods for the social sciences*. Boston: Allyn and Bacon.
- Lofland, J., & Lofland, L. H. (1995). *Analyzing social settings: A guide to qualitative observation and analysis*. Belmont, CA: Wadsworth.
- Taylor, S. J., & Bogdan, R. (1998). *Introduction to qualitative research methods: A guide & resource*. New York: Wiley.

FILE DRAWER PROBLEM

The file drawer problem represents a threat to the VALIDITY of the results and conclusions of a META-ANALYSIS or research synthesis. This phenomenon, also known as *publication bias*, or occasionally as *retrieval bias*, reflects a concern resulting from the fact that meta-analysts generally are not able to retrieve and include the entire POPULATION of studies that were conducted on the topic of interest. The term itself was coined by Robert Rosenthal (1979) to suggest the possibility that the studies that remained tucked away in researchers' file drawers contained results that were different from those included in the meta-analysis, and could therefore potentially invalidate the results of the meta-analysis.

The file drawer problem is based on the belief that there is a SELECTION BIAS operating such that studies overestimating the magnitude of an effect are more likely to find their way into a meta-analysis than are studies underestimating the magnitude of the effect. This is because, for any given sample size, the study that overestimates the TREATMENT effect is more likely to have a significant *P* VALUE and is thus more likely to be published. Because published studies are easier to locate and retrieve than unpublished ones, the possibility arises that a meta-analysis based primarily on published studies may be unrepresentative of the body of research as a whole.

Three kinds of techniques have been developed to help meta-analysts deal with publication bias. One set of techniques is designed to detect publication bias and is best exemplified by a graphical diagnostic called the funnel plot (Light & Pillemer, 1984), although it also includes explicit statistical tests such as those

proposed by Mattias Egger and his colleagues (Egger, Smith, Schneider, & Minder, 1997). The second set of techniques is usually referred to as file drawer analyses because it typically involves imputing the number of EFFECT SIZE estimates with zero effects (corresponding to the unpublished studies left in researchers' file drawers) that would be necessary to reduce the observed meta-analytic result to zero, or, alternatively, to a theoretically or clinically meaningless level (Orwin, 1983). These are based on Rosenthal's (1979) original suggestion of computing the number of "missing" studies with zero effect that would be necessary to reduce the meta-analytic result to statistical insignificance. This number has come to be known as the Fail-safe *N*. When the required number of "missing" studies is implausibly large, the meta-analytic findings may be considered robust to the potential effects of publication bias. Meta-analyses based on small numbers of studies are particularly vulnerable to the file drawer problem. In this case, a small number of unretrieved studies with results that differ from those that were retrieved and included could weaken or overturn the meta-analytic findings.

The third set of techniques is designed to adjust effect size estimates for the possible effects of publication bias under some explicit model of publication selection. The best known of these techniques is the trim-and-fill method developed by Sue Duvall and Richard Tweedie (2000).

Prospectively, the file drawer problem may be ameliorated by a comprehensive search strategy that includes unpublished studies, by research registries, and by changes in editorial policies.

—Hannah R. Rothstein

REFERENCES

- Duvall, S., & Tweedie, R. (2000). Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 276-284.
- Egger, M., Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629-634.
- Light, R., & Pillemer, D. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
- Orwin, R. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics*, 8, 157-159.
- Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86, 638-641.

FILM AND VIDEO IN RESEARCH

Since the early 20th century, film and, more recently, video have been used across the social sciences to visually record and represent human behavior and culture. Film was first used in QUALITATIVE RESEARCH at the end of the 19th century, when, for example, the anthropologist A. C. Haddon used film in an expedition to the Torres Straits. Anthropologists produced film as archive reference footage or as filmic representations of culture, developing the documentary genre of *ethnographic film*. Robert Flaherty's film *Nanook of the North* (1921), a dramatic depiction of the lives of North American Eskimos, is known as the first ethnographic film. Today, ETHNOGRAPHIC film usually follows an observational film style. Examples include the late-20th-century Granada Television *Disappearing World* series. More recently, technological developments in digital video that allow higher image quality and less costly editing processes have attracted ethnographic filmmakers to video rather than film. For example, David MacDougall (2001) has argued that digital video allows ethnographers to produce films that correspond better to academic principles, in contrast to the high budget films that conformed to the demands of TV broadcast standards. MacDougall discusses how his use of digital video in his research in India has allowed him to explore new issues and questions in greater depth and to collaborate more closely with the subjects of his films.

The use of film and video as a research method also continued throughout the 20th century. Film and video continued to be used to visually record research activities such as FOCUS GROUPS, and to study body movement and gestures. This has included extending the use of video in PARTICIPANT OBSERVATION and IN-DEPTH INTERVIEWING (both already aspects of ethnographic filmmaking). For example, in research with a dance group, Thomas (1997) participated by videotaping rehearsals, which the group later viewed. Barnes, Taylor-Brown, and Weiner (1997) videotaped spoken statements made by HIV-positive mothers for their children. Like Thomas, Barnes et al. also produced visual materials that were useful to their sociological analysis but also had a purpose for the subjects of their research. In this case, the videos were to be viewed by the women's children in the future. Similarly, researchers might ask informants to produce their own video diaries of aspects of their everyday

lives, experiences, and identities. Such methods allow researchers to produce both a visual and an oral record of an interview or research situation, as well as a visual text that represents the experience of research and that can therefore be interrogated to understand the processes through which knowledge was produced in research.

Developments in digital video technology have made video research methods increasingly popular in ethnographic research since the late 1990s. Indeed, the possibilities offered by digital video suggest that it will become a popular visual research tool. New digital technologies also allow video to be transferred directly from a camera to a personal computer. This facilitates either digital video editing of ethnographic films or the digital analysis and storage of video and other research materials. Thus, with limited technical skills, contemporary visual researchers can work relatively independently.

—Sarah Pink

REFERENCES

- Barnes, D. B., Taylor-Brown, S., & Weiner, L. (1997). "I didn't leave y'all on purpose": HIV-infected mothers' videotaped legacies for their children. *Qualitative Sociology*, 20(1), 7–32.
- MacDougall, D. (2001). Renewing ethnographic film: Is digital video changing the genre? *Anthropology Today*, 17(3), 15–21.
- Pink, S. (2001). *Doing visual ethnography: Images, media and representation in research*. London: Sage.
- Thomas, H. (1997). Dancing: Representation and difference. In J. McGuigan (Ed.), *Cultural methodologies*. London: Sage.

FIRST-ORDER. See ORDER

FISHING EXPEDITION

Fishing expedition describes a situation in which the researcher has no particular aims for the research at hand. Generally, the term may refer (often in a negative way) to situations where researchers do not form hypotheses based on prior research but simply search for results to discuss as if on a "fishing expedition." More specifically, it may also describes situations in which statistical data are crunched without any

particular guiding hypotheses in the hope of finding statistically significant relationships because, if one uses the .05 level of significance, there is a 1 in 20 chance that a significant result will be found.

—Tim Futing Liao

FIXED EFFECTS MODEL

ANALYSIS OF VARIANCE (ANOVA) models can be used to study how a dependent variable Y depends on one or several factors. A factor here is defined as a categorical independent variable; in other words, an explanatory variable with a nominal LEVEL OF MEASUREMENT. A fixed effects ANOVA model (also called a Type I ANOVA model) has fixed (i.e., nonrandom) parameters expressing the effect of the categories of each factor. Such a model is appropriate if the statistical inference aims at finding conclusions that hold for precisely the categories of the factors that are present in the data set at hand. This contrasts with RANDOM-EFFECTS MODELS, which are described in another article and can be appropriate if the observations of the categories of the factor are a sample from a population. An example of a fixed effects ANOVA model could be a study about work satisfaction of employees depending on gender and a classification in a small number of job categories.

Fixed effects ANOVA models are special cases of the GENERAL LINEAR MODEL, which is basic to REGRESSION analysis. The effects of the categories of a factor in an ANOVA can be expressed as regression coefficients of dummy variables in a regression analysis. The general linear model can be formulated as

$$Y_i = b_0 + b_1X_{1i} + \cdots + b_pX_{pi} + E_i,$$

where i indicates the case, Y is the dependent variable, X_1 to X_p are the independent (or explanatory) variables, and E is the unexplained part, usually called the residual or error term. The quantity b_h is the regression coefficient of variable X_h . The regression coefficients are fixed (i.e., nonrandom) quantities and also are called fixed effects. They are characteristics of the population for which the regression model is defined. Fixed effects occur similarly in other, more complicated models, such as GENERALIZED LINEAR MODELS or nonlinear regression models.

There is another reason for modeling effects of a given factor as fixed rather than random, irrespective

of whether the categories of the factor can be regarded as a sample from a population. This is for factors or other variables that are not of primary interest for the researcher but that have to be included in the model to achieve a good model specification; such variables are called *nuisance variables* or *control variables*. In many cases (depending on the correct model specification), such nuisance variables are better controlled for by including them with fixed effects than with random effects. “Better control” means that the parameters of primary interest can be estimated with less bias, or without bias altogether. One class of examples is panel data, used in econometrics, where, in some studies, the time variable (often the year date) has the role of a nuisance variable, which is controlled for by including it as a fixed effect (see Greene, 2003). Another example is the study about work satisfaction of employees depending on gender and job categories mentioned earlier, now conducted in a (small or large) number of different companies. The companies, if regarded as a nuisance variable, could be treated as a fixed effect. On the other hand, if there is interest in the amount and kind of variability between the companies, they could be included as a random effect, which would lead to a MIXED-EFFECTS MODEL.

—Tom A. B. Snijders

REFERENCES

- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Greene, W. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Iversen, G. R., & Norpoth, H. (1987). *Analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–001, 2nd ed.). Newbury Park, CA: Sage.
- Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. W. (1996). *Applied linear regression models*. New York: Irwin.

FLOOR EFFECT

A floor effect occurs when a measure possesses a distinct lower limit for potential responses and a large concentration of participants score at or near this limit (the opposite of a CEILING EFFECT). Scale attenuation is a methodological problem that occurs whenever variance is restricted in this manner. The problem is commonly discussed in the context of experimental

research, but it could threaten the validity of any type of research in which an outcome measure demonstrates almost no variation at the lower end of its potential range.

Floor effects can be caused by a variety of factors. If the outcome measure involves a performance task with a lower limit (such as number of correct responses), participants may find the task too difficult and make no correct responses. Rating scales also possess lower limits, and participants' responses may fall outside the lower range of the response scale.

Due to the lack of variance, floor effects pose a serious threat to the validity of both experimental and nonexperimental studies, and any study containing a floor effect should be interpreted with caution. If a dependent variable suffers from a floor effect, the experimental or quasi-experimental manipulation will appear to have no effect. For example, a common assessment of memory is percent of items recalled. If the memory task is too difficult, all participants may demonstrate no recall of items. A study involving the effect of sensory modality on recall may produce null findings, when, in fact, the floor effect will not allow differences to be found.

Floor effects are especially troublesome when interpreting INTERACTION EFFECTS. Many interaction effects involve failure to find a significant difference at one level of an independent variable and a significant difference at another. Failure to find a significant difference may be due to a floor effect, and such interactions should be interpreted with caution.

Floor effects are also a threat for nonexperimental or correlational research. Most inferential statistics are based on the assumption of a normal distribution of the variables involved. It may not be possible to compute a significance test, or the restricted range of variance may increase the likelihood of rejecting the null hypothesis when it is actually true (TYPE II ERROR).

The best solution to the problem of floor effects is pilot testing, which allows the problem to be identified early. If a floor effect is found, the problem may be addressed several ways. If the outcome measure is task performance, the task can be made easier in order to increase the range of potential responses. For rating scales, the number of potential responses can be expanded. An additional solution involves the use of extreme anchor stimuli in rating scales. Prior to assessing the variable of interest, participants can be asked to rate a stimulus that will require them to use the extreme ends of the distribution. This will encourage

participants to use the middle range of the rating scale, decreasing the likelihood of a floor effect.

—Robert M. Hessling, Tara J. Schmidt,
and Nicole M. Traxel

FOCUS GROUP

A focus group is a research interviewing process specifically designed to uncover insights from a small group of subjects. The group interview is distinctive in that it uses a set of questions deliberately sequenced or focused to move the discussion toward concepts of interest to the researcher. The focus group consists of a limited number of homogeneous participants, discussing a predetermined topic, within a permissive and nonthreatening environment. A skilled moderator or interviewer guides the discussion, exercising limited control over the discussion and moving it from one question to another.

Although it is called a focus group interview, it is really more like a focused discussion. The moderator engages the participants in a conversation, and participants are encouraged to direct their comments to others in the group. The primary responsibilities of the moderator are to keep the discussion on topic and on time, and to engage the participants in the conversation.

Focus group interviews are used in a variety of ways and have gained popularity because the procedure has helped researchers understand behaviors, customs, and insights of the target audience. This information has been used to understand how customers respond to new programs or products as well as factors influencing consumer behavior. More recently, focus groups have moved beyond consumer research to include academic research, program evaluation, social marketing, needs assessments, and related studies. Focus groups are increasingly used in research studies to complement and augment other methods, such as the development of a SURVEY instrument or the discovery of alternative interpretations.

DEVELOPMENT OF FOCUS GROUP INTERVIEWING

The term *focus group interview* is derived from Robert Merton's research just before World War II. Merton and Paul Lazarsfeld, sociologists at Columbia University, introduced group interviews that had a

focused questioning sequence that elicited comments about behavior, attitudes, and opinions. Merton, along with Patricia Kendall and Marjorie Fisk, wrote about their studies on propaganda materials and troop morale in *The Focused Interview* (1956/1990).

Focus groups gained popularity during the late 1950s as market researchers sought to understand consumer purchasing behavior. From the 1950s through the 1980s, focus groups were used mostly to uncover consumer psychological motivations. Many prominent focus group practitioners were trained as psychologists and sought to uncover the unconscious connections to consumer behavior. At this time, focus groups were composed of 10 to 12 participants, often consisting of strangers, and conducted behind one-way mirrors to allow the sponsor opportunity to observe.

Since the 1980s, academic researchers have used focus groups as a type of qualitative research. Academics used the foundation work of Merton and his colleagues and adapted procedures used by market researchers. These more recent groups place greater emphasis on conducting the group in a natural environment (homes, public meeting spaces); with fewer participants (often 6 to 8 people); and with additional rigor in the analysis (systematic and verifiable procedures).

Focus group research is an evolving field of study. Over the years, researchers have adapted and changed the techniques to fit particular situations. Telephone or Internet focus groups are gaining in popularity as ways to involve people who are separated by distance. Focus groups have shown considerable versatility in international environments where literacy can impede other forms of inquiry.

QUALITY OF FOCUS GROUP STUDIES

As focus groups grew in popularity, the quality began to suffer. To a casual observer, focus groups can look deceptively easy, but in fact, they are complicated to plan, conduct, and analyze. When researchers conducted focus groups without adequate preparation, planning, or skills, the results were less dependable. Much of the quality of the focus group depends on four key factors:

- Planning, which includes identifying the right audience and getting those individuals to gather together
- Asking the right questions

- Moderating the discussion without leading the conversation into predetermined answers
- Using systematic and verifiable analysis strategies in preparing the report

Planning a focus group study consists of determining the number of groups, developing a strategy for recruitment, creating incentives, and anticipating other logistic concerns. The basic design strategy is to conduct 3 or 4 focus groups for each audience category that is of interest to the researcher. A statistical formula is not appropriate for determining sample size. Instead, researchers use the concepts called *redundancy* or *THEORETICAL SATURATION*. With redundancy or theoretical saturation, the researcher continues interviewing until no new insights are presented. In effect, the researcher has heard the range of views on the topic. Continuing the interviews would only yield more of what the researcher already knows. In focus group research, this tends to occur regularly after 3 or 4 groups with one audience.

Recruiting begins by identifying as precisely as possible the characteristics of the target audience. Participants are selected and invited because they have certain experiences or qualities in common. People are invited who meet these “screens” or qualifications of the study.

There are three distinct qualities of successful recruiting. First, the process should be personalized. This means that each invited person feels that he or she has been personally asked to attend and share his or her opinions. Second, the invitation process is repetitive. Invitations are given not just once, but two or three times. Third, an incentive is used to encourage participation. Incentives could include money, food, gifts, or intangibles, such as the opportunity to help a worthy research effort, assist a neighborhood organization, or further a cause to which the individual is committed. The incentive may vary from one study to another because of the varying interests of the participants.

Questions must be developed for the focus group. These questions are designed to be conversational and easy for the participants to understand. In a 2-hour focus group, there would be approximately a dozen questions. Care is needed in the phrasing and sequencing of the questions. Focus group questions are distinctive in their focus, their emphasis on concrete experiences, and their tendency to elicit conversation.

Here are factors often considered when developing questions:

- Use OPEN-ENDED QUESTIONS to allow the participant as much latitude as possible in responding to the question.
- Avoid questions that can be answered with a “yes” or “no” because these answers do not provide the detailed explanation often needed for the study.
- “Why?” is rarely asked because “why” questions seem demanding and make people defensive. Instead, ask about attributes and/or influences.
- Use “think back” questions to take people back to a specific time to get information based on actual experience.
- Use questions that get participants involved, such as comparing, rating, sorting, drawing, and so on.
- Focus the questions, sequencing them from general to specific so that the questions move seamlessly toward the most important topics.
- Use ending questions to bring closure to the group and get summary comments from participants.

Moderating is the process of guiding the discussion. It begins with helping people feel comfortable enough to share what they think and how they feel in the group. The moderator establishes a trusting, permissive, and comfortable environment that removes barriers to communication. The moderator introduces the study and helps the participants get acquainted with each other. At the beginning of the focus group, the moderator describes the **PROTOCOL** of the discussion and indicates that the conversation will be recorded. The participants are assured of confidentiality and told that although quotes might be used, no names will be attached in later reports. After setting up a few ground rules to guide the discussion, the moderator begins to ask the questions. The moderator faces several concerns as the conversation continues. The moderator must know when to wait for more information and when to move on to the next question. The moderator must be able to control dominant speakers and encourage hesitant participants. The moderator must also respect the participants, listen to what they have to say, and thank them for their views, even when he or she may personally disagree. The moderator guides the participants

through the questions, carefully monitoring the time and moving from question to question when sufficient responses have been provided.

Analysis is the process of making sense out of the series of conversations. Not all studies require the same level of analysis. Occasionally, analysis can be done quickly because the data are clear and patterns are evident. But other times, the analytic process can be extremely time-consuming. The analysis process consists of reviewing the data as captured by multiple means (memory, field notes, tape recording, and/or video recording) and looking for the major themes that cut across groups. Often, the focus group conversation is transcribed, and this transcript is used in later analysis.

Focus group analysis should be systematic and verifiable. Being systematic means that the analyst has a protocol that follows a predetermined sequence for capturing and handling the data. The process is verifiable in that there is a trail of evidence that could be replicated.

WHEN TO USE FOCUS GROUPS

As with all research methods, there are situations where focus groups can be particularly effective and situations where other methods should be used.

Focus groups work particularly well for the following:

- Discovering what prompts certain behavior, such as purchasing, using, or taking action
- Discovering the barriers that impede certain behaviors
- Pilot testing ideas, campaigns, surveys, products, and so on. Focus groups can be used to get reactions to plans before big amounts of money are spent in implementation.
- Understanding how people think or feel about something, like an idea, a behavior, a product, or a service
- Developing other research instruments, such as surveys or **CASE STUDIES**
- Evaluating how programs or products are working and how they might be improved
- Understanding how people see needs and assets in their lives and communities

In contrast, focus group interviews are not meant to be used as

- A process for getting people to come to consensus.
- A way to teach knowledge or skills.
- A test of knowledge or skills.

Focus groups do not work well when

- The purpose is to make statistical inferences to a population. The numbers involved in focus group research are too small for statistical projections.
- The topic is so highly controversial that participants want to argue instead of discuss.
- The environment does not foster trust and respect. If the sponsoring agency or organization is not considered trustworthy, or if events in the community make people hesitant or cynical, then the study results will be jeopardized.
- The topic is not appropriate for group discussion and should be discussed individually. Note: Although the researcher has promised confidentiality, in reality, he or she cannot guarantee that the participants will maintain confidentiality.

APPROACHES TO FOCUS GROUP INTERVIEWING

The focus group interview is an evolving methodology that has been adapted to different situations and environments. What follows are some of the most common forms.

Market Research Focus Group Approach

In these groups, professional moderators conduct groups with 10 to 12 homogeneous participants who are paid to participate on a topic relating to a consumer product, service delivery, or organizational identity. These groups are often conducted in special focus group rooms with one-way mirrors, around an oval table for a 2-hour period. Results and reports are rarely shared with participants. The participants typically do not know the identity of the sponsoring organization. Often, the results are needed within weeks.

Academic Focus Group Approach

In these groups, the moderator is an academic researcher who conducts groups of 6 to 8 homogeneous

participants on a topic of scholarly interest that might provide insight into a theoretical concept or an understanding of the reality of the participants. These groups are usually held in convenient and comfortable community settings instead of special rooms with one-way mirrors. Participants might receive a financial incentive for participating. The results are later published in scholarly journals or in dissertations. Considerable emphasis is placed on careful data collection and systematic analytical processes. These studies are typically conducted over a period of 6 to 12 months.

Public/Nonprofit Focus Group Approach

In these groups, the moderator could be an internal staff person, a volunteer, or an outside consultant. The groups are smaller (often 6 to 8 participants) than the market research approach. The groups are conducted in community settings, and one-way mirrors usually are avoided because of the rental cost and also the sensitivity to being observed. The topics are designed to improve services or products, evaluate programs and policies, understand customer satisfaction, or evaluate other needs of public, religious, educational, or service organizations. Almost always, the results are shared back with community members. These studies are often conducted over a period of several months.

Participatory Focus Group Approach

These focus groups are led by volunteers, community members, or possibly staff who are willing to work together as a research team. Often, a professional researcher serves as a mentor and coach, helping to prepare the team, develop the questions, and take on the complicated task of analysis. The volunteer researchers go into their communities, recruit participants, and conduct focus groups using predetermined questions. These studies have been effective in gaining insight into prevention efforts, health and safety concerns, and community needs assessments. The primary advantage of these groups is that the volunteers have a rapport within the community and can quickly establish trust and candor in the conversations. These results are shared back with community members and are often gathered over a period of several months.

Telephone Focus Group Approach

The telephone focus group is similar to a conference phone call. It is led by a moderator and includes 4 to 6 participants who are linked together by telephone. The primary advantage is the convenience. A researcher can conduct a group with people scattered across the nation at a fraction of the cost of having them together in person. Usually, fewer questions (5–8) are sent out in advance of the group. These groups often last about 1 hour.

Children Focus Groups

These groups are modified to fit the interests and abilities of the children. The participants are close in age (usually less than 2 years apart) because of developmental differences. Participants could be as young as 10 years of age, but they must be able to listen to others in the group and carry on a conversation. The group size is small (5–7), and the groups may be a bit shorter in length. The questions usually involve more activities and exercises than what is present in an adult group.

Internet Focus Groups and Chat Rooms

These groups are evolving and have considerable variation. Some groups are similar to chat lines with a small number of participants who are logged on to a central chat room. The moderator poses questions online, and the participants offer their answers in writing. Another variation occurs when the participants log on to a special Web site and receive a different question each day for a series of days. The questions often build on each other, and participants can see the comments of other participants. As technology advances, there will undoubtedly be additional options for these electronic focus groups.

—Richard A. Krueger

REFERENCES

- Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research* (3rd ed.). Thousand Oaks, CA: Sage. [This one-volume (215 page) overview of focus group interviewing provides a description of what focus groups are, how they are used, and strategies that make them successful.]
- Merton, R. K., Fiske, M., & Kendall, P. L. (1990). *The focused interview* (2nd ed.). New York: Free Press. (Originally

published in 1956, Merton and his colleagues describe the development of the focused interview.)

Morgan, D. L., & Krueger, R. A. (Eds.). (1998). *Focus group kit*. Thousand Oaks, CA: Sage. (A collection of six books, each approximately 100–140 pages, on focus group interviewing. Specific volumes include: *The focus group guidebook*, *Planning focus groups*, *Developing questions for focus groups*, *Moderating focus groups*, *Involving community members in focus groups*, and *Analyzing and reporting focus group results*.)

FOCUSED COMPARISONS. See STRUCTURED, FOCUSED COMPARISON

FOCUSED INTERVIEWS

Most commonly known as the FOCUS GROUP method, this form of interview was originally formulated for use with both individuals and groups (Merton, 1987; Merton & Kendall, 1946). The explicit objective of the focused interview is to test, appraise, or produce hypotheses about a particular concrete situation in which the respondent(s) have been involved (e.g., a shared event or salient experience). The *focus* of the interview is circumscribed by relevant theory and evidence and involves skilled facilitation of the process (in a one-on-one or group forum) using an INTERVIEW GUIDE, allowing for unanticipated views to also be uncovered and explored. Optimal use of the method involves an appreciation of the paradox involved in balancing the quest for authentic subjective information through free-flowing discussion with the need for methodological rigor.

The purpose of the focused interview is to go beyond the summary judgments made by respondents about their experiences (e.g., unpleasant, stimulating) to “discover precisely what further feelings were called into play” (Merton & Kendall, 1946, p. 541). Hence, a key to the technique is the skilled use of verbal prompts to activate a concrete response to questions rather than vague generalizations, coupled with an ability to continually monitor and adjust the interview process to maximize the likelihood of self-disclosure. Applied to groups, the process facilitation challenge includes harnessing the synergy potential of multiple respondents to ensure that both a

breadth and depth of evidence is derived (Millward, 2000). Important design considerations in this instance include group size (e.g., large enough to elicit a broad range of views while not producing inhibitions), group composition (e.g., educational homogeneity), and spatial arrangements (e.g., circular seating) (Merton, 1987).

The focused one-on-one interview is now an important part of the methodological repertoire, especially in clinical and health research contexts. It was not until the late 1980s, however, that the research potential of the focused group interview was formally acknowledged by social scientists, beyond its use in marketing quarters (Morgan, 1988). Both types of focused interview can be used either as a primary means of qualitative data collection (e.g., for hypothesis formulation) or as an adjunct to other methods (e.g., to aid interpretation of findings). Theory can also be used as a focusing vehicle affording constructs around which to anchor either one-on-one or group dialogue (e.g., protection motivation theory), particularly in association with complex or sensitive issues (e.g., examination of risky adolescent sexual behavior).

Depending on respondent consent, interview responses are audiotaped or videotaped and then transcribed in preparation for some form of qualitative data analysis. At its most basic level, CONTENT ANALYSIS involves inducing categories of content. Other, more specialized forms of analysis applicable to focused interview data include, for example, interpretative phenomenological analysis and DISCOURSE ANALYSIS. The latter examines both the process and the content of discussion. Ultimately, choice of analytic strategy will depend on a combination of both epistemological and theoretical factors, including precisely what question is being asked (Millward, 2000).

—Lynne Millward

REFERENCES

- Greenbaum, T. L. (1998). *The handbook for focus group research*. London: Sage.
- Merton, R. K. (1987). The focused interview and focus groups: Continuities and discontinuities. *Public Opinion Quarterly*, 51, 550–566.
- Merton, R. K., Fiske, M., & Kendall, P. L. (1990). *The focused interview: A manual of problems and procedures* (2nd ed.). New York: Free Press.
- Merton, R. K., & Kendall, P. L. (1946). The focused interview. *American Journal of Sociology*, 51(6), 514–557.
- Millward, L. J. (2000). Focus groups. In G. M. Breakwell, C. Fife-Schaw, & S. Hammond (Eds.), *Research methods in psychology* (2nd ed., pp. 303–324). London: Sage.
- Morgan, D. L. (1988). *Focus groups as qualitative research*. Newbury Park, CA: Sage.
- Morgan, D. L. (1993). *Successful focus groups: Advancing the state of the art*. London: Sage.
- Stewart, D. W., & Shamdasani, P. N. (1990). *Focus groups: Theory and practice* (Applied Social Research Methods Series, Vol. 20). Newbury Park, CA: Sage.

FORCED-CHOICE TESTING

Forced-choice testing, in which test takers must choose among alternatives, is distinguished from free-response testing, in which test takers provide a response of their own construction. In the social sciences, choosing between these formats is generally dictated by the goals of testing; forced-choice testing lends itself to rapid assessment over a broad content area, resulting in data readily quantifiable for interpretation. Free-response testing is directed toward qualitative analysis of response and generally requires much more time to interpret responses.

Forced-choice testing comprises a variety of formats. Multiple-choice tests present the test taker with two or more choices for response. Two-alternative, forced-choice tests include “true-false,” “yes-no,” or “correct-incorrect” answer pairs. Multiple-choice ability tests include one correct response; the remainder are incorrect alternatives. Attitudinal measures include LIKERT-type formats in which the responder chooses a measure of agreement with a statement, typically from a numerically weighted five-point spread ranging from “strongly agree” to “strongly disagree”; the midpoint choice is typically “neutral” or “undecided.” A Q-sort technique is useful for attitudinal and self-concept measurement. Individuals sort attitudinal statements or personality traits into groups ranging from least characteristic to most characteristic of themselves. Typically, the distribution of groups is forced, preventing a skewing of responses.

Personality tests have adopted strategies of assessing the consistency of responding (whether similarly worded items are answered in similar fashion), the presence of random responding (“content

nonresponsive”), and under- or overreporting of psychopathology (Greene, 2000). In certain examination settings (e.g., personnel selection and child custody evaluations), test takers may be motivated to underreport psychopathology. In pretrial criminal examinations, personal injury litigation, and disability determinations, test takers may be motivated to overreport psychopathology. Simple counts compared to normative data allow quick determination of response sets.

In personality testing, random responding can be assessed by consistency-of-response scales. In ability testing, completely random responding usually is not at issue except when the test taker perceives an advantage in being seen as impaired. Forced-choice testing allows an analysis of the contribution of guessing toward the total number of correct responses. A good estimate of the true number of items known can be computed by $[(\text{number correct}) - (\text{number incorrect})]/(n - 1)$, where n is the number of alternatives to choose between or among (Cronbach, 1990).

Forced-choice testing enables systematic analysis of feigned impairment on cognitive and sensory tests. Pankratz (1979) popularized the technique of comparing the number of correct responses to that expected by random responding; scoring “below chance” indicates malingering. Frederick and Crosby (2000) showed that Cronbach’s technique could result in an estimate of true cognitive capacity for overt malingering. Furthermore, Frederick and Crosby demonstrated that plotting average response accuracy against average response difficulty for two-alternative, forced-choice tests with a hierarchy of difficulty allowed classification of responding as compliant, inconsistent, random, or “suppressed” (malingered).

—Richard I. Frederick

REFERENCES

- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: HarperCollins.
- Frederick, R. I., & Crosby, R. D. (2000). Development and validation of the Validity Indicator Profile. *Law and Human Behavior*, 24, 59–82.
- Greene, R. L. (2000). *The MMPI-2: An interpretive manual*. Boston: Allyn & Bacon.
- Pankratz, L. (1979). Symptom validity testing and symptom retraining: Procedures for the assessment and treatment of functional sensory deficits. *Journal of Consulting and Clinical Psychology*, 47, 409–410.

FORECASTING

It is far better to foresee even without certainty than not to foresee at all.

Henri Poincaré, *The Foundations of Science*, 1913, p. 129

A forecast is a statement about an unknown and uncertain event—most often, but not always, a future event. If time is involved, a forecast (or prediction) is an assertion about a future outcome that is based on observed regularities among events in the past. We assume that observations on a single time series become available at consecutive time periods. The observations may be monthly sales data, quarterly GNP figures, yearly homicides, or daily stock price closings. Past observations $y_1, y_2, \dots, y_n$ are used to calculate a forecast $\hat{y}_n(r)$ of the future observation y_{n+r} . Here, n is the forecast origin, r is the forecast lead time, and $\hat{y}_n(r)$ is the r -step-ahead forecast from time origin n . The question is, how should the available observations be used when calculating the forecast? Here, we assume that there are no other series that can help with the forecast of future y s.

Exponential SMOOTHING and forecasting with ARIMA (autoregressive integrated moving average) time series models represent two successful approaches to univariate time series forecasting. We describe these methods in detail and point to useful extensions.

EXPONENTIAL SMOOTHING PROCEDURES

One can distinguish three basic procedures: simple exponential smoothing for series without trend, Holt’s exponential smoothing for data exhibiting trend, and Winters exponential smoothing for data that include trend and seasonal components.

Simple Exponential Smoothing

An intuitive approach to the prediction of a series without trend is to use the historic average as prediction for all future periods. That is,

$$\hat{y}_n(r) = \bar{y} = \frac{1}{n}[y_n + y_{n-1} + \dots + y_1].$$

Equal weighting of the observations, irrespective of their age, requires stability of the series over time.

Because most data series lack this stability, it is better to either average observations over a short, recent time window, or consider a weighted average that discounts observations according to their age. Simple exponential smoothing uses an exponential (geometric) weighted average:

$$\hat{y}_n(r) = L_n = \alpha[y_n + (1 - \alpha)y_{n-1} + (1 - \alpha)^2 y_{n-2} + (1 - \alpha)^3 y_{n-3} + \dots].$$

L_n is the estimate of the level at time n . It becomes the forecast of all future observations. The smoothing constant α ($0 \leq \alpha \leq 1$) gives more weight to recent observations and less weight to observations in the past. When $\alpha = 1$, the forecast $\hat{y}_n(r) = L_n = y_n$ makes use of the last observation only. It is called the “naïve” or “RANDOM WALK” forecast. When α approaches 0, the resulting forecast $\hat{y}_n(r) = L_n = (y_n + y_{n-1} + \dots + y_1)/n$ puts equal weight on all observations.

The level is updated according to $L_n = \alpha y_n + (1 - \alpha)L_{n-1}$; only the last observation and the previous level estimate need to be stored. This is important if many series must be predicted. It is common to start the recursion at the beginning of the series, with a certain starting value for the level at Time 1. Typically, one uses the first observation ($L_1 = y_1$), or an average of the first few observations. From this starting value, one calculates $L_2 = \alpha y_2 + (1 - \alpha)L_1$, $L_3 = \alpha y_3 + (1 - \alpha)L_2$, ..., until one reaches $L_n = \alpha y_n + (1 - \alpha)L_{n-1}$. The last level estimate becomes the forecast of all future observations: $\hat{y}_n(r) = L_n$.

The calculations require a value for the smoothing constant. The smoothing constant is estimated by

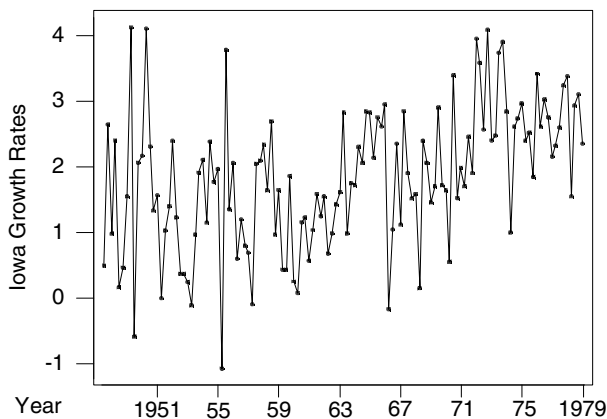


Figure 1 Quarterly Iowa Growth Rates (1948 through 1979)

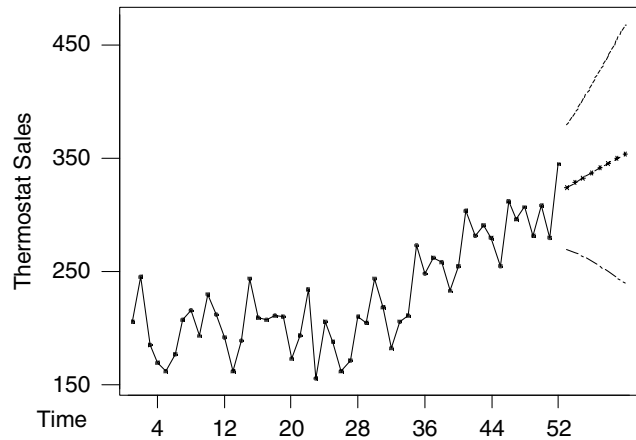


Figure 2 Observations, Forecasts, and 95% Forecast Intervals

minimizing the sum of the squared one-step-ahead forecast errors,

$$\text{SSE} = \sum_{t=2}^n [y_t - \hat{y}_{t-1}(1)]^2 = \sum_{t=2}^n (y_t - L_{t-1})^2.$$

We illustrate simple exponential smoothing with the quarterly Iowa nonfarm growth rates in Figure 1. The data ($n = 127$; from 1948/2 through 1979/4) were first analyzed in Abraham and Ledolter (1983). The time series graph shows no obvious trends, but indicates that the level of the series shifts with time. Table 1 illustrates the calculations with smoothing constant $\alpha = 0.1$. The iterations start with $L_1 = y_1 = 0.50$. Repeated use of the updating equation gives $L_2 = (0.1)(2.65) + (0.9)(0.5) = 0.715$, $L_3 = (0.1)(0.97) + (0.9)(0.715) = 0.741$, until $L_{127} = (0.1)(2.35) + (0.9)(2.681) = 2.648$. Forecasts of all future observations are given by $\hat{y}_{127}(r) = 2.648$. The one-step-ahead forecast errors are given in the fifth column: $y_2 - \hat{y}_1(1) = 2.65 - 0.50 = 2.15$, $y_3 - \hat{y}_2(1) = 0.97 - 0.715 = 0.255$, ..., $y_{127} - \hat{y}_{126}(1) = 2.35 - 2.681 = -0.331$; the sum of their squares is $\text{SSE} = 122.61$. The sum of squares depends on the smoothing constant, because a different α implies different forecasts. In this example, any smoothing constant different from 0.1 increases the sum of the squared one-step-ahead forecast errors.

Holt's Exponential Smoothing

Holt extends the exponential smoothing approach to the situation where the series follows a linear

Table 1 Simple Exponential Smoothing for Iowa Growth Rates

<i>Time</i>	<i>Growth Rate</i>	<i>Smoothed Value</i>	<i>One-Step-Ahead Forecast</i>	<i>One-Step-Ahead Forecast Error</i>
1	$y_1 = 0.50$	$L_1 = 0.50$		
2	$y_2 = 2.65$	$L_2 = 0.715$	$\hat{y}_1(1) = 0.50$	$y_2 - \hat{y}_1(1) = 2.15$
3	$y_3 = 0.97$	$L_3 = 0.741$	$\hat{y}_2(1) = 0.715$	$y_3 - \hat{y}_2(1) = 0.255$
4	$y_4 = 2.40$	$L_4 = 0.906$	$\hat{y}_3(1) = 0.741$	$y_4 - \hat{y}_3(1) = 1.659$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
124	$y_{124} = 1.55$	$L_{124} = 2.602$	$\hat{y}_{123}(1) = 2.719$	$y_{124} - \hat{y}_{123}(1) = -1.169$
125	$y_{125} = 2.93$	$L_{125} = 2.635$	$\hat{y}_{124}(1) = 2.602$	$y_{125} - \hat{y}_{124}(1) = 0.328$
126	$y_{126} = 3.10$	$L_{126} = 2.681$	$\hat{y}_{125}(1) = 2.635$	$y_{126} - \hat{y}_{125}(1) = 0.465$
127	$y_{127} = 2.35$	$L_{127} = 2.648$	$\hat{y}_{126}(1) = 2.681$	$y_{127} - \hat{y}_{126}(1) = -0.331$

trend. The r -step-ahead forecast of y_{n+r} is calculated from

$$\hat{y}_n(r) = L_n + rT_n,$$

where L_n and T_n reflect the level and the trend of the series at time n . Estimates of these quantities are needed, and update recursions need to be specified. Holt considers the recursions

$$\begin{aligned} L_n &= \alpha_1 y_n + (1 - \alpha_1)(L_{n-1} + T_{n-1}), \\ T_n &= \alpha_2(L_n - L_{n-1}) + (1 - \alpha_2)T_{n-1}. \end{aligned}$$

The two smoothing constants, α_1 and α_2 , reflect the weights that are given to current (“new”) and previously available (“old”) information. The most recent information on the level at time n is given by y_n . Prior to this observation, we have available L_{n-1} and T_{n-1} and our best guess of the next level is $L_{n-1} + T_{n-1}$. The updating equation for the level specifies a weighted average of these two components. The second equation for the trend specifies a weighted average of the most recent trend estimate, $L_n - L_{n-1}$, and the previously available estimate T_{n-1} .

Large smoothing constants put more weight on recent observations. Smoothing constants $\alpha_1 = \alpha_2 = 1$ imply $L_n = y_n$, $T_n = y_n - y_{n-1}$, and $\hat{y}_n(r) = y_n + r(y_n - y_{n-1})$. All forecasts lie on a straight line through the last two observations. For small smoothing constants, the observations are weighted equally, similar to a regression of the observations on time.

Updating calculations starts with reasonable initial values such as $L_2 = y_2$ and $T_2 = y_2 - y_1$. The recursions are used to compute L_3 and T_3 , L_4 and

$T_4, \dots$, until L_n and T_n are reached. Forecasts follow from $\hat{y}_n(r) = L_n + rT_n$. Smoothing constants are estimated by minimizing the sum of the squared one-step-ahead forecast errors.

Holt’s method is applied to the thermostat sales data analyzed in Abraham and Ledolter (1983). Figure 2 shows the $n = 52$ observations, the next eight forecasts, and approximate 95% forecast intervals. Minitab is used for the estimation of the smoothing constants and the calculation of the forecasts.

Winters Seasonal Exponential Smoothing

Many series include seasonal components. Sales are typically seasonal, with large sales around Christmas. For seasonal data, one extends Holt’s approach with an additional equation and a third smoothing constant. Assuming monthly data with a yearly seasonal component, the forecasts are given by

$$\begin{aligned} \hat{y}_n(r) &= (L_n + rT_n)S_{n+r-12} \quad \text{for } r = 1, 2, \dots, 12 \\ &= (L_n + rT_n)S_{n+r-24} \quad \text{for } r = 13, \dots, 24 \\ &= \dots \end{aligned}$$

The difference to the forecasts in Holt’s approach is the adjustment of trend forecasts by appropriate seasonal factors. Here, the adjustment is multiplicative, but additive adjustments are also possible.

Update equations for the estimates of level (L_n), trend (T_n), and seasonal components (S_n) are needed.

Again, these equations specify weighted averages of “new” and “old”:

$$L_n = \alpha_1(y_n/S_{n-12}) + (1 - \alpha_1)(L_{n-1} + T_{n-1})$$

$$T_n = \alpha_2(L_n - L_{n-1}) + (1 - \alpha_2)T_{n-1}$$

$$S_n = \alpha_3(y_n/L_n) + (1 - \alpha_3)S_{n-12}.$$

The three smoothing constants, α_1, α_2 , and α_3 , determine how observations are weighted. The first two equations are similar to Holt's equations, with the only difference being that the most recent observation is adjusted for seasonality (y_n/S_{n-12}). The update of the seasonal factor in the last equation is a weighted average of the last available “old” seasonal factor (S_{n-12}) and the “new” information on the seasonal factor, y_n/L_n .

Smoothing constants are obtained by minimizing the sum of the squared one-step-ahead forecast errors. The calculations are started with initial values for the level, trend, and seasonal components; estimates are updated recursively until the last available time period is reached; and forecasts are calculated. The computations are more complicated, but conceptually not different from the simple versions.

Forecast Intervals

Point forecasts are important, but they mean little if one cannot attach a level of CONFIDENCE. An approximate 95% r -step-ahead forecast interval for the future observation y_{n+r} is given by

$$\hat{y}_n(r) \pm 2s_r$$

where the ROOT MEAN SQUARE error

$$s_r = \sqrt{\sum_{t=r+1}^n [y_t - \hat{y}_{t-r}(r)]^2 / (n - r)}$$

is calculated from the “observed” r -step-ahead forecast errors, $y_{r+1} - \hat{y}_1(r), y_{r+2} - \hat{y}_2(r), \dots, y_n - \hat{y}_{n-r}(r)$.

BOX-JENKINS ARIMA MODELS

Exponential smoothing is a procedure that, when applied to a time series, results in forecasts. The BOX-JENKINS ARIMA approach is *model-based*. First, one determines the model for the underlying stochastic process that has generated the series, and then one uses the model to derive the optimal forecasts.

Autoregressive integrated moving average (ARIMA) models view the autocorrelated observations y_t as the

result of a linear filter on uncorrelated random variables a_t . The filter is written as a ratio of polynomials, leading to the difference equation

$$\varphi(B)y_t = \phi(B)(1 - B)^d y_t = \theta_0 + \theta(B)a_t,$$

where

- B is the backshift operator, $B^m y_t = y_{t-m}$
- a_t are uncorrelated, unobserved errors with constant variance
- $\varphi(B) = \phi(B)(1 - B)^d = 1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_{p+d} B^{p+d}$ is the generalized autoregressive operator of order $p + d$
- d is the degree of differencing needed to transform a nonstationary series to stationarity
- $\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ is a stationary autoregressive operator of order p
- $\theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$ is an invertible moving average operator of order q
- θ_0 is a constant that introduces deterministic trends when $d \geq 1$.

Special Cases

The first-order ARIMA(1, 0, 0) model, $(1 - \phi B)y_t = \theta_0 + a_t$ or $y_t = \theta_0 + \phi y_{t-1} + a_t$, is a regression of the current observation on its lag. In the ARIMA($p, 0, 0$) model, the observation at time t is regressed on the p previous lags. The ARIMA(0, 1, 0) random walk model, $(1 - B)y_t = \theta_0 + a_t$ or $y_t = \theta_0 + y_{t-1} + a_t$, is useful for modeling stochastic nonstationarity; the constant θ_0 introduces additional deterministic drift. The ARIMA(0, 1, 1) model, $(1 - B)y_t = (1 - \theta B)a_t$ or $y_t = y_{t-1} + a_t - \theta a_{t-1}$, is another commonly used model. Model extensions to cover seasonal situations are available.

Box and Jenkins (1976) study the properties of these models and outline a useful model-building strategy. The model order (p, d, q) is obtained from the patterns in the AUTOCORRELATIONS and partial autocorrelations of the series, the parameters in the models are estimated by MAXIMUM LIKELIHOOD, and the adequacy of the models is assessed through residual diagnostics.

Forecasts Implied by ARIMA Models

Optimal (minimum mean square error) forecasts of future observations are obtained by taking *conditional expectations* of future observations. These conditional expectations can be derived from the difference

equation of the ARIMA model. We illustrate this on the second-order autoregressive model,

$$y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + a_t.$$

The conditional expectations of future observations ($y_{n+1}, y_{n+2}, \dots$) are

$$\hat{y}_n(1) = \theta_0 + \phi_1 y_n + \phi_2 y_{n-1}$$

$$\hat{y}_n(2) = \theta_0 + \phi_1 \hat{y}_n(1) + \phi_2 y_n$$

$$\hat{y}_n(r) = \theta_0 + \phi_1 \hat{y}_n(r-1) + \phi_2 \hat{y}_n(r-2) \quad \text{for } r \geq 3.$$

Better understanding of the forecasts can be obtained through the *eventual forecast function*. The eventual forecast function of the ARIMA(p, d, q) model is given by

$$\begin{aligned} \hat{y}_n(r) &= f_1(r)\beta_1^{(n)} + f_2(r)\beta_2^{(n)} + \dots + f_{p+d}(r)\beta_{p+d}^{(n)} \\ &= f(r)^{\text{tr}} \beta^{(n)} \quad \text{for } r > [q - (p + d)], \end{aligned}$$

where $f(r) = [f_1(r), f_2(r), \dots, f_{p+d}(r)]^{\text{tr}}$ is a column vector of forecast functions that depend only on the generalized autoregressive operator (tr denotes the vector transpose). For example, in the ARIMA(0, 2, q) model, the forecast functions are given by $f_1(r) = 1$ and $f_2(r) = r$, and the eventual forecast function is linear.

The vector of coefficients $\beta^{(n)} = [\beta_1^{(n)}, \beta_2^{(n)}, \dots, \beta_{p+d}^{(n)}]^{\text{tr}}$ is updated with each new observation according to

$$\beta^{(n)} = L\beta^{(n-1)} + h[y_n - \hat{y}_{n-1}(1)].$$

The transition matrix L is a consequence of the forecast functions $f(r)$, and the vector h depends on the parameters in the ARIMA model.

Relationships Among ARIMA Forecasts and Exponential Smoothing

There are similarities of the eventual forecast functions implied by ARIMA models and the forecast functions in exponential smoothing. Autoregressive components and the degree of differencing imply the forecast functions of the ARIMA forecasts. Updates of the coefficients in the eventual forecast function and in exponential smoothing can be made the same if certain special ARIMA models are selected.

ARIMA models are more general. Forecasts from constrained ARIMA models coincide with exponential smoothing forecasts. Forecasts from the ARIMA

(0, 1, 1) model, $(1 - B)y_t = (1 - \theta B)a_t$, coincide with those from simple exponential smoothing provided that $\theta = 1 - \alpha$. Forecasts from the ARIMA(0, 2, 2) model, $(1 - B)^2 y_t = (1 - \theta_1 B - \theta_2 B^2)a_t$, are the same as the forecasts in Holt's exponential smoothing provided that $\theta_1 = 2 - \alpha_1(1 + \alpha_2)$ and $\theta_2 = -(1 - \alpha_1)$. Relationships among the seasonal forecast procedures are more complicated (see Abraham & Ledolter, 1986).

Despite these theoretical results, forecast competitions show that in many practical applications, simple exponential smoothing procedures outperform the more general ARIMA models (see Makridakis & Hibon, 2000). A lack of robustness of more complicated and in all likelihood incorrect models, and problems with fitting nonparsimonious models with unnecessary parameters can explain such findings.

IMPORTANT ISSUES NOT COVERED IN THIS REVIEW

Many other univariate forecast procedures have been developed but are not covered in this review. The structural time series models described in Harvey (1990) and the stochastic models developed for finance applications (see Tsay, 2002) should be mentioned. Our review deals with quantitative methods but ignores judgmental forecast approaches that are important if little or no data are available (see Armstrong, 2001).

Univariate time series forecasts use only the series's own past history. One would expect that forecasts of Y are improved when incorporating the information from other related series X . Transfer function models can be used if y_t is affected by current or lagged x s, and forecast improvements will result if there is a strong lead-lag relationship. A contemporaneous association of y_t with x_t alone is usually of little benefit, because one needs a forecast for X in order to predict Y . Multivariate time series models, particularly multiple autoregressive models, are appropriate if feedback among the series is present. ECONOMETRIC models use systems of many equations to represent the relationships among economic variables. The purpose of econometric models is to provide conditional, or "what-if," forecasts, as compared to the unconditional forecasts provided by univariate forecast procedures.

Many different forecast methods are available. Each method has its strengths and weaknesses, and relies on ASSUMPTIONS that are often violated. Theoretical and empirical evidence suggests that a combination

of forecasts is advantageous (*International Journal of Forecasting*, No. 4, 1989).

COMPUTER SOFTWARE

Excellent software is available for time series forecasting. All major statistical software packages (such as SAS, SPSS, and Minitab) include useful time series forecasting modules. In addition, several special-purpose programs (such as ForecastPro and SCA) include “automatic” procedures that can identify the most appropriate techniques without much user intervention. This is important in large-scale forecast applications where one must predict thousands of series.

SUGGESTED READINGS

The books by Abraham and Ledolter (1983), Newbold and Bos (1994), and Diebold (2001) give a good introduction to this area. *Principles of Forecasting*, edited by Armstrong (2001), includes summaries and references to commonly used forecast methods and covers both quantitative and qualitative (judgmental) forecast procedures. The *Journal of Forecasting* and the *International Journal of Forecasting* are two journals specializing in forecasting issues.

—Johannes Ledolter

REFERENCES

- Abraham, B., & Ledolter, J. (1983). *Statistical methods for forecasting*. New York: Wiley.
- Abraham, B., & Ledolter, J. (1986). Forecast functions implied by ARIMA models and other related forecast procedures. *International Statistical Review*, 54, 51–66.
- Armstrong, J. S. (Ed.). (2001). *Principles of forecasting*. Boston: Kluwer.
- Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control* (rev. ed.). San Francisco: Holden-Day.
- Diebold, F. X. (2001). *Elements of forecasting* (2nd ed.). Cincinnati, OH: South-Western.
- Harvey, A. C. (1990). *Forecasting, structural time series models, and the Kalman filter*. New York: Cambridge University Press.
- Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions, and implications. *International Journal of Forecasting*, 16, 451–476.
- Newbold, P., & Bos, T. (1994). *Introductory business forecasting* (2nd ed.). Cincinnati, OH: South-Western.
- Tsay, R. S. (2002). *Analysis of financial time series*. New York: Wiley.

FORESHADOWING

The term is used by some qualitative researchers to refer to the process of specifying issues in which one is interested in advance of entering the field. As the British social anthropologist Bronislaw Malinowski wrote: “Preconceived ideas are pernicious in any scientific work, but foreshadowed problems are the main endowment of the scientific thinker, and these problems are first revealed to the observer by his theoretical studies” (Malinowski, 1922, pp. 8–9, cited in Hammersley & Atkinson, 1995, pp. 24–25). Thus, unlike the open-ended approach that is sometimes advocated by qualitative researchers, Malinowski and others who have followed him recommended coming to the field with an awareness of certain issues that require investigation.

—Alan Bryman

REFERENCES

- Hammersley, M., & Atkinson, P. (1995). *Ethnography: Principles in practice*. London: Routledge.
- Malinowski, B. (1922). *Argonauts of the Western Pacific*. London: Routledge & Kegan Paul.

FOUCAULDIAN DISCOURSE ANALYSIS

Michel Foucault developed a complex and nuanced theory of discourse in a series of books that includes *The History of Sexuality* (1998) and *The Archeology of Knowledge* (2002). Although his work covered a broad array of empirical topics and substantial shifts in method, three concepts recur throughout his work and provide some coherence to his varied interests: power, knowledge, and subjectivity. Understanding Foucauldian discourse analysis begins with an understanding of Foucault’s CONCEPTUALIZATION of these three concepts and their relationship.

Foucault defines discourses, or discursive formations, as bodies of knowledge that form the objects of which they speak. In other words, discourses do not simply describe the social world; they constitute it by bringing certain phenomena into being through the way in which they categorize and make sense of an otherwise meaningless reality. The nature of the

discursive formation in place at any point in time is the source of the relations of power (and resistance), the social objects and identities, and the possibilities for speaking and acting that exist at any point in time.

Each discourse is defined by a set of rules or principles—the “rules of formation”—that lead to the appearance of particular objects that make up recognizable social worlds. Discourse lays down the “conditions of possibility” that determine what can be said, by whom, and when. No statement occurs accidentally, and the task for the discourse analyst is to analyze how and why one particular statement appeared rather than another.

For Foucault, discourse—or at least the knowledge that it instantiates—is inseparable from power. Power is embedded in knowledge, and any knowledge system constitutes a system of power, as succinctly summarized in his “power/knowledge” couplet. In constructing the available identities, ideas, and social objects, the context of power is formed. Power, in turn, brings into being new forms of knowledge and produces new social objects.

Power is not something connected to agents but represents a complex web of relations determined by systems of knowledge constituted in discourse. It is thus the discursive context, rather than the subjectivity of any individual actor, that influences the nature of political strategy. In fact, for Foucault, the notion of agents acting purposefully in some way not determined by the discourse is antithetical.

In determining the conditions of possibility in this way, discourse “disciplines” subjects in that actors are known—and know themselves—only within the confines of a particular discursive context. Thus, discourse shapes individuals’ subjectivity. Foucault decenters the subject: There is no essence or true subject but only a power/knowledge system that constituted the subjectivities of the individuals embedded within it.

Foucault’s work has had a tremendous influence on social studies and, in particular, on understandings of the relation between discourse and power. At the same time, his empirical method is highly idiosyncratic and has found only limited footing as yet beyond his own work at an empirical level. However, his work is highly influential and stands to make an even more significant contribution to social science in the future.

—Nelson Phillips

REFERENCES

- Foucault, M. (1998). *The history of sexuality: The will to knowledge* (Vol. 1). London: Penguin.
- Foucault, M. (2002). *Archeology of knowledge*. London: Routledge.

FRAUD IN RESEARCH

Fraud in the natural sciences (especially the biological and medical) is not uncommon (Broad & Wade, 1982). Fraud in social scientific research, however, is said to be comparatively rare. But of course, it all depends on what is meant by “fraud.” This is a contested concept, both in general and in particular cases. This inherent uncertainty has two sources: the shifting and fuzzy character of activities that may come under this description, and the “paradox of success.”

Typologies of research fraud abound. Charles Babbage’s 1830 description of the illegitimacies of “trimming,” “cooking,” and “forging” (Broad & Wade, 1982, pp. 29–30) is an early and often-cited example of a typology with an “action-type” criterion. The alternative kind uses an “intention-type” criterion to distinguish, for example, the fraud from the hoax, where the former is perpetrated with serious and permanent intent, and the latter is done as a joke and is intended to be temporary. Thus, the perpetrator of fraud is treated with greater opprobrium than the hoaxer.

The intentional fraud is a paradoxical object, however. Because any instance can be known only through its discovery, and because its discovery achieves its failure, *successful* frauds cannot be known to exist—or not to exist. For some commentators, like Robert Joynson, one of Cyril Burt’s later defenders (see below), this can be problematic: “‘The hallmark of a *successful* confidence-trickster is precisely that people are persuaded of his honesty.’ If this strange argument is accepted, everyone who is trusted must be regarded as a *successful* confidence-trickster” (Joynson, 1989, p. 302). Strictly, what this “paradox of success” specifies is that honest work and action cannot be distinguished from their fraudulent, and undiscovered, counterparts. The implication of this is that it is vital to understand the work of its discovery (debunking) if we wish to understand the phenomenon of fraud (Ashmore, 1995).

What accounts for the relative rarity of full-blooded fraud in social research? One reason is that

opportunities for its *discovery* are relatively scarce, and the resources required for such work are consequently large. Although the biologist John Darsee can be, apparently, observed “one evening in May 1981 flagrantly [forging] raw data” (Broad & Wade, 1982, p. 14), no such simple debunking techniques are available to convict a cultural anthropologist like Margaret Mead (see “Special Section,” 1983) or even a quantitative psychologist like Cyril Burt. Although Burt was indeed accused, by Leon Kamin and others, of flagrantly forging raw data, this debunking effort was unstable—vulnerable to later “meta-debunking” work that accused the accusers of equivalent crimes (Joynson, 1989; see Mackintosh, 1995, for the current “not proven” consensus). The more qualitative and discursive the style of social research, the less likely it is that fraud, with the exception of plagiarism, will be discoverable.

—Malcolm Ashmore

REFERENCES

- Ashmore, M. (1995). Fraud by numbers: Quantitative rhetoric in the Piltdown forgery discovery. *South Atlantic Quarterly*, 94(2), 591–618.
- Broad, W., & Wade, N. (1982). *Betrayers of the truth: Fraud and deceit in the halls of science*. New York: Simon & Schuster.
- Joynson, R. B. (1989). *The Burt affair*. London: Routledge.
- Mackintosh, N. J. (1995). *Cyril Burt: Fraud or framed?* Oxford, UK: Oxford University Press.
- Special section: Speaking in the name of the real: Freeman and Mead on Samoa. (1983). *American Anthropologist*, 85, 908–947.

FREE ASSOCIATION INTERVIEWING

The Free Association Narrative Interview (FANI) is a qualitative method for the production and analysis of interview data that combines a narrative emphasis with the psychoanalytic principle of free association. Questions that elicit narratives encourage interviewees to remember specific events in their experiences. These, unlike generalized accounts, possess emotional resonance. The principle of free association is based on the idea that it is the unconsciously motivated links between ideas, rather than just their contents, that provide insight into the emotional meanings of interviewees' accounts (see PSYCHOANALYTIC METHODS).

Therefore, it is particularly appropriate for exploring questions about interviewees' identities or that touch emotions and sensibilities (in contrast to methods designed to elicit opinions or facts).

In most empirical social science methods, research subjects are assumed to know themselves well enough to give reliable accounts. The FANI method instead posits defenses against anxiety as a core feature of research subjects, as of all subjects (including researchers). Because anxiety characterizes the research relationship and will be more or less prominent depending on the topic, the setting, and the degree of rapport, defenses against anxiety will potentially compromise interviewees' ability to know the meaning of their actions, purposes, and relations. Such unconscious defenses are one feature of a psychosocial subject, part of the psychic dimension produced by a unique biography structured by defenses, desires, and mental conflict. These influence and are influenced by the social dimension, which consists of a shared social world that is understood as a combination of intersubjective processes, real events, and the discourses through which events are rendered meaningful.

The method was developed as an adaptation of the biographical interpretative method (see BIOGRAPHIC NARRATIVE INTERPRETIVE METHOD) in the context of a project investigating the relationship between anxiety and fear of crime (Hollway & Jefferson, 2000). A narrative question format about specific events (“Can you tell me about a time when . . .?”) was adopted in order to discourage responses of a generalized and rationalized kind, as frequently elicited by a SEMISTRUCTURED INTERVIEW. OPEN-ENDED questions were asked, and “why” questions were avoided. Interviewees' answers were followed up in their order of telling, using their phrasing. These protocols are designed to keep within participants' own meaning frames rather than the interviewer's.

The FANI method, using gestalt principles, entails holding the whole data set in mind when interpreting the meaning of an extract. Consistent with psychoanalytic principles, rather than expect coherence, the researcher is alert to signs of incoherence and conflict in the text, such as changes in emotional tone, contradictions, and avoidances.

During interviews and data analysis, interviewers use their own feelings as data, following psychoanalytic principles of transference and countertransference, in order to identify their own emotional investments. Each participant is interviewed twice.

Two researchers listen to the first audiotape before the second interview, and the process of noticing the significance of links, associations, and contradictions contributes to the formulation of narrative questions for the second interview.

—Wendy Hollway and
Tony Jefferson

REFERENCE

Hollway, W., & Jefferson, T. (2000). *Doing qualitative research differently: Free association, narrative and the interview method*. London: Sage.

FREQUENCY DISTRIBUTION

Frequency distribution shows the number of observations falling into each of the categories or ranges of values of a VARIABLE. Frequency distributions are often portrayed by way of a HISTOGRAM, a (FREQUENCY) POLYGON, or a frequency table. When such distribution is expressed in terms of proportions or percentages, it is known as RELATIVE FREQUENCY DISTRIBUTION. When the distribution represents incremental frequencies, it is known as cumulative frequency distribution.

—Tim Futing Liao

FREQUENCY POLYGON

A frequency polygon is one of several types of graphs used to describe a relatively large set of quantitative data. An example using a set of adult male heights is given in Figure 1.

To create a frequency polygon, data are first sorted from high to low and then grouped into contiguous intervals. The intervals are represented on the x -axis of a graph by their midpoints. The y -axis provides a measure of frequency. It shows the number of data points falling in each interval. It is scaled from a minimum of 0 to a maximum that exceeds the frequency of the modal interval. Within the field of the figure, straight lines connect points that mark the frequencies for each interval. These lines should extend to the x -axis, which

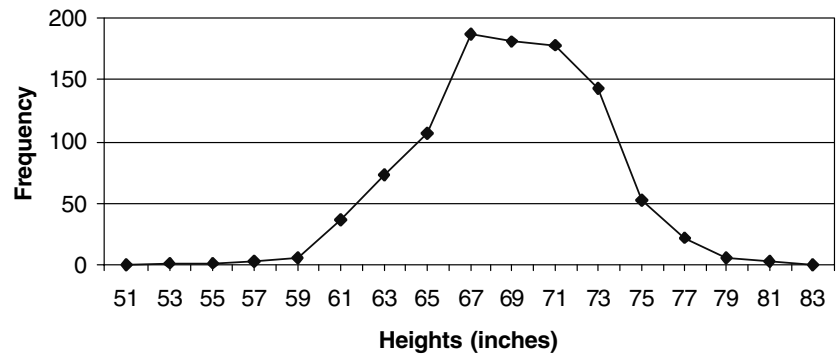


Figure 1 Adult Male Heights Displayed According to the Frequency for Each 2-Inch Interval Between 52 and 82 Inches

means that empty intervals need to be included at the low and high end of the data set.

Frequency polygons are used to represent quantitative data when the data intervals do not equal the number of distinct values in the data set. This is always the case for continuous data. When the data are discrete, it is possible to represent every value in the dataset on the x -axis. In this case, a HISTOGRAM would be used instead of a frequency polygon. The importance of the distinction is that points are connected with lines in the polygon to allow interpolation of frequencies for x -axis values not at the midpoints. When data have no values other than those represented on the x -axis, frequencies are simply marked by bars (or columns).

If the y -axis in a frequency polygon is used to denote proportions rather than frequencies, the graph is known as a relative frequency polygon. For some data displays, it is more useful to show cumulative frequencies for data intervals rather than absolute frequencies. In this case, a cumulative frequency polygon (or ogive) would be used to describe the data set.

—Kenneth O. McGraw

FRIEDMAN TEST

The Friedman (1937) test provides a test for the NULL HYPOTHESIS of equal TREATMENT effects in a two-way layout. This EXPERIMENTAL DESIGN is also known as the balanced complete BLOCK DESIGN, randomized complete block design, and REPEATED-MEASURES DESIGN.

Suppose we have t treatments and b blocks and a total of N experimental units, where N is a multiple

of the product bt . For convenience, we assume that $N = bt$ so that there are exactly t experimental units within each block. The t treatments are RANDOMLY ASSIGNED to the t units within the i th block to obtain observations $X_{i1}, X_{i2}, \dots, X_{it}$. To carry out the test, these observations in the i th block are ranked from smallest to largest using the integer ranks $1, 2, \dots, t$. This is repeated for each block. Let R_{ij} denote the rank assigned to the experimental unit in block i that received treatment j . The complete set of ranked data can then be presented in the following two-way table:

Block	Treatment				Sum
	1	2	...	t	
1	R_{11}	R_{12}	...	R_{1t}	$t(t+1)/2$
2	R_{21}	R_{22}	...	R_{2t}	$t(t+1)/2$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
b	R_{b1}	R_{b2}	...	R_{bt}	$t(t+1)/2$
Sum	R_1	R_2	...	R_t	N

We note that the row sums are each equal to $1 + 2 + \dots + t = t(t+1)/2$. If the treatment effects are, indeed, all the same, as under the null hypothesis, the ranks in each row are simply a random permutation of $1, 2, \dots, t$, and the column sums will tend to be equal. But the total of the column sums must equal $N = bt(t+1)/2$, and therefore, each column sum must be equal to $bt(t+1)/2$ divided by t , or $b(t+1)/2$. The test statistic S is the sum of squares of the deviations between the actual column sums and the sum expected under the null hypothesis, or

$$\begin{aligned}
 S &= \sum_{i=1}^t [R_i - b(t+1)/2]^2 \\
 &= \sum_{i=1}^t R_i^2 - b^2 t(t+1)^2/4.
 \end{aligned}$$

If there is a definite ordering among the treatment effects, all of the ranks in some column will be equal to 1, all of the ranks in some other column will be equal to 2, and so on, and the column sums will be some permutation of the numbers $1b, 2b, \dots, tb$. It can

Table 1 Data on Currency Fluctuations

Currency	Monday	Tuesday	Wednesday	Thursday	Friday
AD	2	3	5	4	1
BP	2	3	4	5	1
JY	1	4	3	5	2
EE	1	3	4	5	2
Total	6	13	16	19	6

be shown that these column sums produce the largest possible value of S . Therefore, the null hypothesis of equal treatment effects should be rejected in favor of the alternative of a definite ordering among treatments for large values of S .

The exact null distribution of S has been tabled for small values of b and t and is reproduced in Gibbons (1993, 1997). For larger sample sizes, the test statistic is $Q = 12S/bt(t+1)$, which is approximately chi-square distributed with $t-1$ DEGREES OF FREEDOM.

As a numerical example, suppose we want to compare median daily fluctuations in foreign currency exchange relative to the U.S. dollar for the Australian dollar (AD), British pound (BP), Japanese yen (JY) and European euro (EE). The fluctuations are measured by a skewness coefficient calculated for each day of the week. The relative values are shown in Table 1, where rank 1 = smallest value. We have $t = 5$ treatments (the days of the week) and $b = 4$ blocks (the currencies). We calculate $S = 138$ and $Q = 13.8$ with 4 degrees of freedom. The approximate p value is $p < .01$, and we reject the null hypothesis of equal median fluctuations. It appears that there is a significant day effect, with the largest fluctuations occurring in the middle of the week and the smallest fluctuations at the beginning and end of the week.

—Jean D. Gibbons

REFERENCES

- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 675-701.
- Gibbons, J. D. (1993). *Nonparametric statistics: An introduction*. Newbury Park, CA: Sage.
- Gibbons, J. D. (1997). *Nonparametric methods for quantitative analysis* (3rd ed.). Columbus, OH: American Sciences Press.

FUNCTION

Scientific accounts of the way a particular phenomenon or process works typically begin by identifying the fundamental actors and the fundamental quantities involved—that is, the variables and the entities to which those variables are attached. Immediately, the challenge is to specify the relations among the fundamental quantities. Many of these relations are summarized and embodied in mathematical functions of the form

$$y = f(x_1, \dots, x_n),$$

where y and the x s are the fundamental quantities, now termed the DEPENDENT and INDEPENDENT VARIABLES, respectively.

As one begins to think about the relations among the quantities, the steps include thinking about which variables operate as inputs for which outcomes; and restricting attention to one function, whether the outcome increases or decreases as each input increases; and whether the rate of increase or decrease is constant, increasing, or decreasing. Of course, one also thinks about the domain and the range of the function, elements that are closely linked to the mathematical representation of the fundamental quantities. Moreover, sometimes the relation between an input and an outcome is nonmonotonic, with the effect changing direction at some point.

Increased understanding about the way a particular process works is manifested in progress from no function to a general function to a specific function, as it becomes clear that some variables belong in the equation (as x s) and others do not; that y increases with some x s, decreases with others, and is nonmonotonically related to still others; and that the rates of change are different with respect to different x s. Typically, the theoretical and empirical literature, even at the prefunction stage, is rich with insights about these features. One approach to specifying the function is to identify in the literature conditions that a function representing the particular relation should satisfy. Indeed, even when a specific function appears on the scene via intuition or empirical work, it is important to strengthen its foundation by building arguments that link its features to the desiderata in the literature.

EXAMPLE

The study of distributive-retributive justice identifies two actors—an *observer* and a *rewardee* (who may, but need not, be the observer)—and three quantities—the rewardee's *actual reward*; the observer's idea of the *just reward* for the rewardee; and the observer's judgment that the rewardee is justly or unjustly rewarded and, if unjustly rewarded, whether overrewarded or underrewarded and to what degree, called the *justice evaluation*. The literature provided many insights for specifying the justice evaluation function. First, the literature made it clear that (for goods) the justice evaluation increases with the actual reward and decreases with the just reward. Second, the literature included a representation of the justice evaluation by real numbers, with zero representing the point of perfect justice and the negative and positive numbers representing, respectively, underreward and overreward. Third, the literature also suggested two desirable properties for a justice evaluation function: (a) scale invariance, so that the justice evaluation is independent of the reward's units of measurement; and (b) additivity, in the sense that the effect of the actual reward on the justice evaluation is independent of the level of the just reward, and conversely. The first two insights led to a general function, and the third to a proof that, in the case of cardinal rewards, one functional form uniquely satisfies both the scale invariance and additivity conditions (Jasso, 1990).

The discipline of reasoning about a relation in order to specify the function has a high payoff. Even when it does not lead to a unique functional form, it serves to clarify scientific understanding of the way the process works.

The functions widely used in social science include psychophysical functions in psychology; utility, production, and consumption functions in economics; earnings functions in economics and sociology; perception and status functions in sociology and social psychology. As knowledge grows, more and more relations become specified as functions. Homans's (1967, p. 18) vivid challenge to specify the function is always timely.

Functions can serve as the starting ASSUMPTIONS for theories and theoretical models; they can also be used directly in empirical work (Jasso, in press). Thus, functions lend coherence to topical domains.

FURTHER READING

Regular perusal of functions in mathematical handbooks (such as Bronshtein and Semendyaev, 1979/1985) and calculus texts (such as Courant, 1934/1988) girds the mind for confrontation with vague relations (and provides immediate pleasure). Reading about the famous functions in social science and their development, such as Fechner's (1860, 1877) work on the sensation function and Mincer's (1958, 1974) on the earnings function, does the same.

—Guillermina Jasso

REFERENCES

- Bronshtein, I. N., & Semendyaev, K. A. (1985). *Handbook of mathematics* (K. A. Hirsch, Trans. & Ed.). New York: Van Nostrand Reinhold. (Originally published in 1979.)
- Courant, R. (1988). *Differential and integral calculus* (2 vols.) (E. J. McShane, Trans.). New York: Wiley. (Originally published in 1934.)
- Fechner, G. T. (1860). *Elemente der psychophysik*. Leipzig: Breitkopf & Härtel.
- Fechner, G. T. (1877). *In sachen der psychophysik*. Leipzig: Breitkopf & Härtel.
- Homans, G. C. (1967). *The nature of social science*. New York: Harcourt, Brace, & World.
- Jasso, G. (1990). Methods for the theoretical and empirical analysis of comparison processes. *Sociological Methodology*, 20, 369–419.
- Jasso, G. (in press). The tripartite structure of social science analysis. *Sociological Theory*.
- Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy*, 66, 281–302.
- Mincer, J. (1974). *Schooling, experience and earnings*. New York: Columbia University Press.

FUZZY SET THEORY

Fuzzy set theory deals with sets or categories whose boundaries are blurry or “fuzzy.” For instance, an object is not necessarily just red or not red, it can be “reddish” or even a “warm green” (green with a tinge of red in it). Likewise, a political scientist may rate one government as more or less democratic than another. Fuzzy set theory permits membership in the set of red objects or democratic governments to be a

matter of degree. It offers an analytic framework for handling concepts that are simultaneously categorical and dimensional. Below, the basic ideas behind fuzzy set theory and fuzzy logic are introduced and, where possible, illustrated with examples from the social sciences.

CONCEPT OF A FUZZY SET

Membership in a set may be represented by values in the $[0,1]$ interval. A classical *crisp* set permits only the values 0 (nonmembership) and 1 (full membership). Treating the set of males as a crisp set would entail assigning each person a 1 if the criteria for being a male were satisfied and a 0 otherwise. A *fuzzy* set, on the other hand, permits values in between 0 and 1. Thus, a 21-year-old might be accorded full membership in the set of “young adults,” but a 28-year-old might be given a degree of membership somewhere between 0 and 1. Conventionally, $1/2$ denotes an element that is neither completely in nor out of the set. A *membership function* maps values or states from one or more *support* variables onto the $[0,1]$ interval. Figure 1 shows two hypothetical membership functions using age as the support: “young adult” and “adolescent.” Although adolescents are generally younger than young adults, these two sets overlap and permit nonzero degrees of membership in each.

Although there is no universally accepted interpretation of what degree of membership in a fuzzy set means, consensus exists about several points. First, a degree of membership is not a probability and, indeed, may be assigned with complete certainty. Moreover, unlike probabilities, degrees of membership need not sum to 1 across an exhaustive set of alternatives. Membership in some contexts is defined in terms of similarity to a prototype, but in others, it may connote a degree of compatibility or possibility in relation to a concept. An example of the latter is “several,” considered to be a fuzzy verbal number. The integer 6 might be considered completely compatible with “several,” whereas 3 would be less so and therefore assigned a lower membership value.

The concept of possibility also has been used to characterize degree of membership. *Possibility* provides an upper envelope on probability. For example, if 45% of the people in a community own a bicycle, then in a random survey of that community, the maximum probability of finding a person who has used his or her bicycle on a given day will be 0.45, but

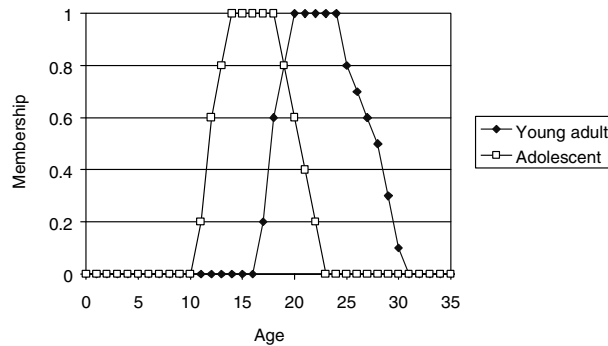


Figure 1 Membership Functions for “Young Adult” and “Adolescent”

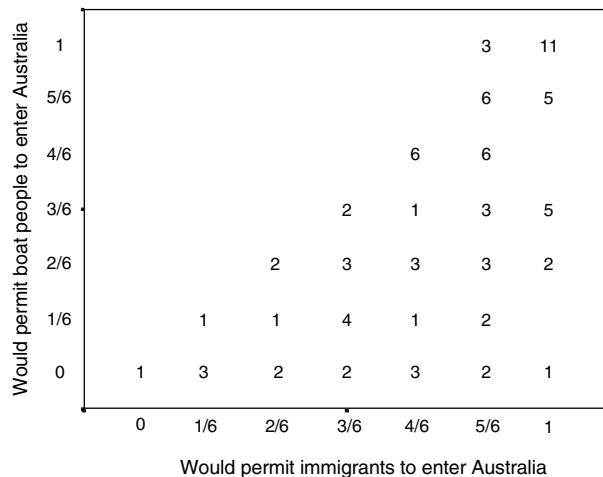


Figure 2 Example of Fuzzy Set Inclusion

of course, the actual probability may well be lower. *Possibility theory* is a framework based on fuzzy set theory and closely related to theories of imprecise probabilities (e.g., lower and upper probabilities and belief functions).

There has been some debate over the meaning of *fuzziness*. It is widely agreed that fuzziness is not ambiguity, vagueness, nonspecificity, unreliability, or disagreement. Instead, fuzziness refers to a precise kind of gradedness. The color judgments of an artist, for instance, are not ambiguous, nonspecific, or vague in any pejorative sense. Various measures of fuzziness have been proposed, ranging from generalizations of information-theoretic entropy to relative variation. Common to all of them is the notion that a set is less fuzzy as more of its elements have membership values

near 0 or 1 and fuzzier as more membership values are near $1/2$.

The measurement level of a membership function also has been debated, although important aspects of fuzzy set theory may be used even for ordinal-level membership functions. In many applications, a reasonable case can be made for the endpoints (0 and 1), but it is more difficult to argue for equal-interval intermediate membership values. A useful way of representing membership at the ordinal level is via *level sets*, which are crisp sets nested by ordered cutting points on the membership scale. Six of these are represented in Figure 1 by the horizontal gridlines. At a membership level of 0.6, for example, the set of young adults includes ages 18 to 27, whereas at membership level 1, the set of young adults shrinks to ages 20 to 24. Level sets are closely related to (de)cumulative distribution functions.

Finally, membership functions may be conditioned by states or membership in other sets. An example of the first is the fuzzy distance “close by” when we are on foot versus traveling by automobile. An example of the second is “tall” when applied to men versus women.

BASIC OPERATIONS ON FUZZY SETS

Let $m_A(x)$ denote the degree of membership element x has in set A . The subscript or (x) may be dropped when there is no ambiguity. Let $\sim A$ denote the complement of A . Fuzzy *negation* is the same as it is in PROBABILITY:

$$m_{\sim A} = 1 - m_A.$$

According to the membership functions in Figure 1, the degree of membership for a 20-year-old in the set of adolescents is 0.6, so the degree of membership in the complementary set is $1 - 0.6 = 0.4$.

Membership in the *intersection* of A and B is defined by

$$m_{A \cap B} = \min(m_A, m_B)$$

and membership in the *union* of A and B is defined by

$$m_{A \cup B} = \max(m_A, m_B).$$

In Figure 1, a 20-year-old is both an adolescent and young adult in degree $\min(0.6, 1) = 0.6$. Fuzzy set unions and intersections are noncompensatory. Unlike a weighted linear combination, for instance, if $m_A > m_B$, then an increase in m_A will not compensate

for the impact of a decrease in m_B on membership in $m_{A \cup B}$. A crucial assumption implicit in these definitions is that the membership functions of sets A and B have a complete joint ordering, so that the comparative operations min and max are permissible.

Set B is said to be a subset of (included by) set A if, for all x , $m_A(x) > m_B(x)$. Note that this definition also requires the joint ordering assumption. Fuzzy set *inclusion* is conceptually a generalization of crisp set inclusion and thereby related directly to GUTTMAN and RASCH scaling. Although it is seldom the case that this inequality is perfectly satisfied, real examples may be readily found where it holds to quite a high degree.

Figure 2 shows one such instance, in which 83 survey respondents rated their degree of agreement with the propositions that “Australia should permit immigrants to enter the country” and “Boat people should be allowed to enter Australia and have their claims processed.” The scatterplot shows only three exceptions to the inclusion relationship, in the upper right-hand corner. The set of people agreeing with the second statement is almost completely a subset of those agreeing with the first statement.

The concepts of inclusion and fuzziness have been used to interpret various transformations of membership functions in terms of approximate natural-language hedges such as “very.” A transformation *concentrates* a fuzzy set if it lowers membership values and *dilates* the set if it increases them. The set of “very tall” men would be said to be a concentration of the set of “tall” men, and the original proposal for a concentration operator was $[m_A(x)]^2$. Although such operators have been strongly criticized as inadequate models of verbal hedges, they provide interpretations of certain transformations.

FUZZY LOGIC

Classical Boolean logic is two-valued, with every proposition being either true or false. Multiple-valued logics permit at least a third state: “indeterminate” or “possible.” In Boolean logic, truth values are limited to 0 (false) and 1 (true), and three-valued logics assign a truth value of 1/2 to the third state. Fuzzy logic generalizes this by allowing any truth value in the $[0, 1]$ interval.

The original versions of fuzzy logic use the min and max operators to compute logical conjunctions and disjunctions. Denoting the truth value of propositions

P and Q by t_P and t_Q , the usual logical primitives are defined by

$$\sim P = 1 - t_P,$$

$$P \wedge Q = \min(t_P, t_Q),$$

$$P \vee Q = \max(t_P, t_Q),$$

$$P \Rightarrow Q = \sim P \vee (P \wedge Q) \quad (\text{Maxmin Rule}) \quad \text{or} \\ = \sim P \vee Q \quad (\text{Arithmetic Rule}), \text{ and}$$

$$P \Leftrightarrow Q = (P \Rightarrow Q) \wedge (Q \Rightarrow P).$$

The Maxmin and Arithmetic Rules yield different truth values for $P \Rightarrow Q$ only when $t_P < t_Q$, as Table 1 shows.

The result is that Arithmetic Rule truth values for implication are greater than or equal to Maxmin Rule truth values. Among the more popular alternative fuzzy logics, the Lukasiewicz (L_1) version yields the most generous truth values for implication. In L_1 , the min and max operators are used for $P \wedge Q$ and $P \vee Q$, but $P \Rightarrow Q = \min(1, 1 - t_P + t_Q)$. Thus, in L_1 , the truth value of $P \Rightarrow Q = 1$ whenever $t_P < t_Q$, which agrees with the definition of subethood in the previous section.

As an example, let us return to the example illustrated in Figure 2 and consider $P \Rightarrow Q$ where P = “Would permit boat people to enter Australia” and Q = “Would permit immigrants to enter Australia.” Figure 3 displays the truth values produced by the Maxmin and Arithmetic Rules for the appropriate positions in the scatterplot from Figure 2. It turns out that if we sum the truth values for the 83 cases in Figure 2, the Maxmin Rule gives a total of 68.33, whereas the Arithmetic Rule gives 75.00. L_1 would assign a truth value of 1 to all but the three cases that violate $t_P < t_Q$, assigning them 5/6 instead, for a sum of 82.50.

Fuzzy logical operations may be combined to produce compound implication statements, which, in turn,

Table 1 Truth Tables for Maxmin and Arithmetic Rules

		Maxmin	Arithmetic
t_P	t_Q	$P \Rightarrow Q$	$P \Rightarrow Q$
0	0	1	1
0	1	1	1
1	0	0	0
1	1	1	1
$p >$	q	$\max(1 - p, q)$	$\max(1 - p, q)$
$p <$	q	$\max(1 - p, p)$	$\max(1 - p, q)$

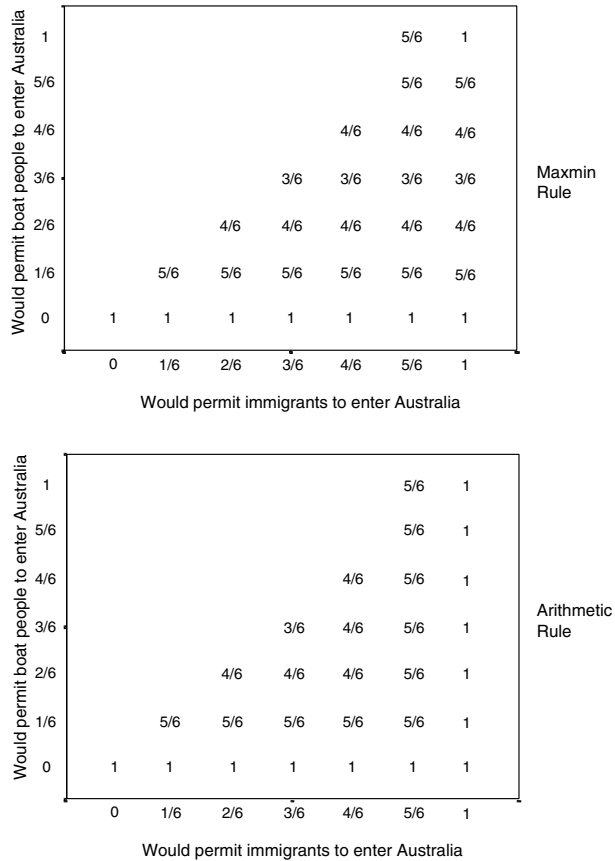


Figure 3 Maxmin and Arithmetic Rule Truth Values

may form the basis for NONLINEAR MODELS. Consider three propositions:

I = It is important to provide information to others.

T = It is important to maintain the trust of a confidant.

C = I am confident that I made the correct decision (to disclose or withhold information).

In an occupational survey, 229 respondents, presented with a dilemma over whether to disclose information confided to them by a colleague, rated each of these statements on 7-point scales that will be treated here as having truth values ranging from 0 to 1. We will examine the fuzzy logical model

$$C \Rightarrow I \vee T.$$

This model proposes that if the decision maker is confident, then she or he rates I or T highly (or both). This is a nonlinear model because the max operator is not linearizable, and, in fact, is not even smooth or differentiable.

The first two subtables in Table 2 show the bivariate distributions of I with C and T with C . The third

subtable shows $I \vee T$ versus C . Visual inspection suggests that in the third subtable, the high-confidence cases have been pulled into the high $I \vee T$ region, whereas the low-confidence cases are fairly evenly scattered along the range of $I \vee T$.

Using the Arithmetic Rule for fuzzy logical implication, the model $C \Rightarrow I$ achieves a total truth value of 159.7 out of 229, and the model $C \Rightarrow T$ achieves 162.3. An averaging model (i.e., $C \Rightarrow [I + T]/2$) does not improve much, obtaining a truth value total of 166.7. However, the truth value of $C \Rightarrow I \vee T$ is 198.00, which is more than a 22% improvement. A somewhat more conventional approach (although not as appropriate) would be to use CORRELATION and REGRESSION to evaluate $I \vee T$ versus a LINEAR model incorporating I and T for predicting C . The correlation between I and C

Table 2 Sufficiency and Joint Necessity

<i>I = Importance of providing information to others</i>								
	0	1/6	2/6	3/6	4/6	5/6	1	Total
1	8	6	4	4	3	7	36	68
5/6		4	12	7	8	28	9	68
4/6		4	4		11	18	5	42
3/6	2	1	9	7	5	8	2	34
2/6				3	2	5	1	11
1/6		1			1			2
0			1	1	1		1	4
Total	10	16	30	22	31	66	54	229

<i>T = Importance of maintaining trust of colleague</i>								
	0	1/6	2/6	3/6	4/6	5/6	1	Total
1	8	4	3	5	4	11	33	68
5/6	1	5	8	9	14	20	11	68
4/6		2	5	9	5	16	5	42
3/6	1			7	8	11	7	34
2/6				2	4	3	2	11
1/6			1			1		2
0	1			1	1		1	4
Total	11	11	17	33	36	62	59	229

<i>I ∨ T</i>								
	0	1/6	2/6	3/6	4/6	5/6	1	Total
1				1	1	13	53	68
5/6			1		11	37	19	68
4/6		1			10	24	7	42
3/6	1			4	7	14	8	34
2/6				2	1	6	2	11
1/6			1			1		2
0				1	2		1	4
Total	1	1	2	8	32	95	90	229

is .114, and between T and C, it is .025. A linear regression model combining both predictors yields a multiple correlation of .126, again not much of an improvement on the largest of the zero-order correlations. However, the correlation between $I \vee T$ and C is .430.

ORIGINS AND DEVELOPMENTS

Fuzzy set theory was invented by Lotfi Zadeh, a professor of engineering at the University of California at Berkeley, and first publicized in his classic 1965 paper. Forerunners include work on multiple-valued logics in the 1920s and 1930s, and Max Black's pioneering 1937 paper on reasoning with vague concepts. The fundamentals of fuzzy logic and possibility theory were worked out in the 1970s. By the late 1980s, axiomatizations of fuzzy set theory and fuzzy logic had been constructed, and debates about their relationship with probability frameworks were well under way, with an emerging consensus becoming apparent toward the end of the 20th century.

Partly because of its origins, the earliest and most extensive applications were made in control engineering. During the 1980s and 1990s, these applications burgeoned into a lucrative industry. In the past 15 years, fuzzy set theory and fuzzy logic have been incorporated along with BAYESIAN probability

and techniques such as NEURAL NETS and genetic algorithms in the "soft computing" approach to constructing intelligent systems.

Applications in the human sciences have been slower to develop. Empirical work in cognitive psychology first appeared the 1970s, along with a few applications in the social sciences. By the end of the 1980s, a number of applications had emerged in subfields ranging from perception and memory to human geography, including at least two books in addition to a sizeable number of journal articles. Although opinion remains divided on the utility of fuzzy set theory for the human sciences, several commentators in recent years have suggested that its potential has not yet been fully realized.

—Michael Smithson

REFERENCES

- Klir, G. J., & Folger, T. A. (1988). *Fuzzy sets, uncertainty, and information*. Englewood Cliffs, NJ: Prentice Hall.
- Smithson, M. (1987). *Fuzzy set analysis for behavioral and social sciences*. New York: Springer-Verlag.
- Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8, 338–353.
- Zetenyi, T. (Ed.). (1988). *Fuzzy sets in psychology*. Amsterdam: North-Holland.

The SAGE Encyclopedia of
**SOCIAL SCIENCE
RESEARCH METHODS**

GENERAL EDITORS

Michael S. Lewis-Beck

Alan Bryman

Tim Futing Liao

EDITORIAL BOARD

Leona S. Aiken

Arizona State University, Tempe

R. Michael Alvarez

California Institute of Technology, Pasadena

Nathaniel Beck

University of California, San Diego, and New York University

Norman Blaikie

*Formerly at the University of Science,
Malaysia and the RMIT University, Australia*

Linda B. Bourque

University of California, Los Angeles

Marilynn B. Brewer

Ohio State University, Columbus

Derek Edwards

Loughborough University, United Kingdom

Floyd J. Fowler, Jr.

University of Massachusetts, Boston

Martyn Hammersley

The Open University, United Kingdom

Guillermina Jasso

New York University

David W. Kaplan

University of Delaware, Newark

Mary Maynard

University of York, United Kingdom

Janice M. Morse

University of Alberta, Canada

Martin Paldam

University of Aarhus, Denmark

Michael Quinn Patton

*The Union Institute & University,
Utilization-Focused Evaluation, Minnesota*

Ken Plummer

University of Essex, United Kingdom

Charles Tilly

Columbia University, New York

Jeroen K. Vermunt

Tilburg University, The Netherlands

Kazuo Yamaguchi

University of Chicago, Illinois

The SAGE Encyclopedia of
**SOCIAL SCIENCE
RESEARCH METHODS**

VOLUME 2

MICHAEL S. LEWIS-BECK

Department of Political Science, University of Iowa

ALAN BRYMAN

Department of Social Sciences, Loughborough University

TIM FUTING LIAO

Department of Sociology, University of Essex and University of Illinois, Urbana-Champaign

EDITORS

A SAGE Reference Publication

 **SAGE Publications**
International Educational and Professional Publisher
Thousand Oaks ■ London ■ New Delhi

Copyright © 2004 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

For information:



SAGE Publications, Inc.
2455, Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
6, Bonhill Street
London EC2A 4PU
United Kingdom

SAGE Publications India Pvt. Ltd.
B-42, Panchsheel Enclave
Post Box 4109
New Delhi 110 017 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Lewis-Beck, Michael S.

The SAGE encyclopedia of social science research methods/Michael S.

Lewis-Beck, Alan Bryman, Tim Futing Liao.

p. cm.

Includes bibliographical references and index.

ISBN 0-7619-2363-2 (cloth)

1. Social sciences—Research—Encyclopedias. 2. Social sciences—Methodology—Encyclopedias.

I. Bryman, Alan. II. Liao, Tim Futing. III. Title.

H62.L456 2004

300!.! 72—dc22

2003015882

03 04 05 06 10 9 8 7 6 5 4 3 2 1

<i>Acquisitions Editor:</i>	Chris Rojek
<i>Publisher:</i>	Rolf A. Janke
<i>Developmental Editor:</i>	Eileen Haddox
<i>Editorial Assistant:</i>	Sara Tauber
<i>Production Editor:</i>	Melanie Birdsall
<i>Copy Editors:</i>	Gillian Dickens, Liann Lech, A. J. Sobczak
<i>Typesetter:</i>	C&M Digital (P) Ltd.
<i>Proofreader:</i>	Mattson Publishing Services, LLC
<i>Indexer:</i>	Sheila Bodell
<i>Cover Designer:</i>	Ravi Balasuriya

List of Entries

Abduction
Access
Accounts
Action Research
Active Interview
Adjusted *R*-Squared. *See* Goodness-of-Fit Measures; Multiple Correlation; *R*-Squared
Adjustment
Aggregation
AIC. *See* Goodness-of-Fit Measures
Algorithm
Alpha, Significance Level of a Test
Alternative Hypothesis
AMOS (Analysis of Moment Structures)
Analysis of Covariance (ANCOVA)
Analysis of Variance (ANOVA)
Analytic Induction
Anonymity
Applied Qualitative Research
Applied Research
Archival Research
Archiving Qualitative Data
ARIMA
Arrow's Impossibility Theorem
Artifacts in Research Process
Artificial Intelligence
Artificial Neural Network. *See* Neural Network
Association
Association Model
Assumptions
Asymmetric Measures
Asymptotic Properties
ATLAS.ti
Attenuation
Attitude Measurement
Attribute
Attrition
Auditing
Authenticity Criteria
Autobiography
Autocorrelation. *See* Serial Correlation
Autoethnography
Autoregression. *See* Serial Correlation
Average
Bar Graph
Baseline
Basic Research
Bayes Factor
Bayes' Theorem, Bayes' Rule
Bayesian Inference
Bayesian Simulation. *See* Markov Chain Monte Carlo Methods
Behavior Coding
Behavioral Sciences
Behaviorism
Bell-Shaped Curve
Bernoulli
Best Linear Unbiased Estimator
Beta
Between-Group Differences
Between-Group Sum of Squares
Bias
BIC. *See* Goodness-of-Fit Measures
Bimodal
Binary
Binomial Distribution
Binomial Test
Biographic Narrative Interpretive Method (BNIM)
Biographical Method. *See* Life History Method
Biplots
Bipolar Scale
Biserial Correlation
Bivariate Analysis
Bivariate Regression. *See* Simple Correlation (Regression)
Block Design
BLUE. *See* Best Linear Unbiased Estimator
BMDP
Bonferroni Technique

- Bootstrapping
Bounded Rationality
Box-and-Whisker Plot. *See* Boxplot
Box-Jenkins Modeling
Boxplot

CAIC. *See* Goodness-of-Fit Measures
Canonical Correlation Analysis
CAPI. *See* Computer-Assisted Data Collection
Capture-Recapture
CAQDAS (Computer-Assisted Qualitative Data Analysis Software)
Carryover Effects
CART (Classification and Regression Trees)
Cartesian Coordinates
Case
Case Control Study
Case Study
CASI. *See* Computer-Assisted Data Collection
Catastrophe Theory
Categorical
Categorical Data Analysis
CATI. *See* Computer-Assisted Data Collection
Causal Mechanisms
Causal Modeling
Causality
CCA. *See* Canonical Correlation Analysis
Ceiling Effect
Cell
Censored Data
Censoring and Truncation
Census
Census Adjustment
Central Limit Theorem
Central Tendency. *See* Measures of Central Tendency
Centroid Method
Ceteris Paribus
CHAID
Change Scores
Chaos Theory
Chi-Square Distribution
Chi-Square Test
Chow Test
Classical Inference
Classification Trees. *See* CART
Clinical Research
Closed Question. *See* Closed-Ended Questions
Closed-Ended Questions
Cluster Analysis
Cluster Sampling
Cochran's Q Test
Code. *See* Coding
Codebook. *See* Coding Frame
Coding
Coding Frame
Coding Qualitative Data
Coefficient
Coefficient of Determination
Cognitive Anthropology
Cohort Analysis
Cointegration
Collinearity. *See* Multicollinearity
Column Association. *See* Association Model
Commonality Analysis
Communality
Comparative Method
Comparative Research
Comparison Group
Competing Risks
Complex Data Sets
Componential Analysis
Computer Simulation
Computer-Assisted Data Collection
Computer-Assisted Personal Interviewing
Concept. *See* Conceptualization, Operationalization, and Measurement
Conceptualization, Operationalization, and Measurement
Conditional Likelihood Ratio Test
Conditional Logit Model
Conditional Maximum Likelihood Estimation
Confidence Interval
Confidentiality
Confirmatory Factor Analysis
Confounding
Consequential Validity
Consistency. *See* Parameter Estimation
Constant
Constant Comparison
Construct
Construct Validity
Constructionism, Social
Constructivism
Content Analysis
Context Effects. *See* Order Effects
Contextual Effects
Contingency Coefficient. *See* Symmetric Measures
Contingency Table
Contingent Effects

-
- Continuous Variable
 Contrast Coding
 Control
 Control Group
 Convenience Sample
 Conversation Analysis
 Correlation
 Correspondence Analysis
 Counterbalancing
 Counterfactual
 Covariance
 Covariance Structure
 Covariate
 Cover Story
 Covert Research
 Cramér's *V*. *See* Symmetric Measures
 Creative Analytical Practice (CAP) Ethnography
 Cressie-Read Statistic
 Criterion Related Validity
 Critical Discourse Analysis
 Critical Ethnography
 Critical Hermeneutics
 Critical Incident Technique
 Critical Pragmatism
 Critical Race Theory
 Critical Realism
 Critical Theory
 Critical Values
 Cronbach's Alpha
 Cross-Cultural Research
 Cross-Lagged
 Cross-Sectional Data
 Cross-Sectional Design
 Cross-Tabulation
 Cumulative Frequency Polygon
 Curvilinearity

 Danger in Research
 Data
 Data Archives
 Data Management
 Debriefing
 Deception
 Decile Range
 Deconstructionism
 Deduction
 Degrees of Freedom
 Deletion
 Delphi Technique
 Demand Characteristics

 Democratic Research and Evaluation
 Demographic Methods
 Dependent Interviewing
 Dependent Observations
 Dependent Variable
 Dependent Variable (in Experimental Research)
 Dependent Variable (in Nonexperimental Research)
 Descriptive Statistics. *See* Univariate Analysis
 Design Effects
 Determinant
 Determination, Coefficient of. *See* *R*-Squared
 Determinism
 Deterministic Model
 Detrending
 Deviant Case Analysis
 Deviation
 Diachronic
 Diary
 Dichotomous Variables
 Difference of Means
 Difference of Proportions
 Difference Scores
 Dimension
 Disaggregation
 Discourse Analysis
 Discrete
 Discriminant Analysis
 Discriminant Validity
 Disordinality
 Dispersion
 Dissimilarity
 Distribution
 Distribution-Free Statistics
 Disturbance Term
 Documents of Life
 Documents, Types of
 Double-Blind Procedure
 Dual Scaling
 Dummy Variable
 Duration Analysis. *See* Event History Analysis
 Durbin-Watson Statistic
 Dyadic Analysis
 Dynamic Modeling. *See* Simulation

 Ecological Fallacy
 Ecological Validity
 Econometrics
 Effect Size
 Effects Coding
 Effects Coefficient

- Efficiency
 Eigenvalues
 Eigenvector. *See* Eigenvalues; Factor Analysis
 Elaboration (Lazarsfeld's Method of)
 Elasticity
 Emic/Etic Distinction
 Emotions Research
 Empiricism
 Empty Cell
 Encoding/Decoding Model
 Endogenous Variable
 Epistemology
 EQS
 Error
 Error Correction Models
 Essentialism
 Estimation
 Estimator
 Eta
 Ethical Codes
 Ethical Principles
 Ethnograph
 Ethnographic Content Analysis
 Ethnographic Realism
 Ethnographic Tales
 Ethnography
 Ethnomethodology
 Ethnostatistics
 Ethogenics
 Ethology
 Evaluation Apprehension
 Evaluation Research
 Event Count Models
 Event History Analysis
 Event Sampling
 Exogenous Variable
 Expectancy Effect. *See* Experimenter Expectancy Effect
 Expected Frequency
 Expected Value
 Experiment
 Experimental Design
 Experimenter Expectancy Effect
 Expert Systems
 Explained Variance. *See* R-Squared
 Explanation
 Explanatory Variable
 Exploratory Data Analysis
 Exploratory Factor Analysis. *See* Factor Analysis
 External Validity
 Extrapolation
 Extreme Case
F Distribution
F Ratio
 Face Validity
 Factor Analysis
 Factorial Design
 Factorial Survey Method (Rossi's Method)
 Fallacy of Composition
 Fallacy of Objectivism
 Fallibilism
 Falsificationism
 Feminist Epistemology
 Feminist Ethnography
 Feminist Research
 Fiction and Research
 Field Experimentation
 Field Relations
 Field Research
 Fieldnotes
 File Drawer Problem
 Film and Video in Research
 First-Order. *See* Order
 Fishing Expedition
 Fixed Effects Model
 Floor Effect
 Focus Group
 Focused Comparisons. *See* Structured, Focused Comparison
 Focused Interviews
 Forced-Choice Testing
 Forecasting
 Foreshadowing
 Foucauldian Discourse Analysis
 Fraud in Research
 Free Association Interviewing
 Frequency Distribution
 Frequency Polygon
 Friedman Test
 Function
 Fuzzy Set Theory
 Game Theory
 Gamma
 Gauss-Markov Theorem. *See* BLUE; OLS
Geisteswissenschaften
 Gender Issues
 General Linear Models
 Generalizability. *See* External Validity

-
- Generalizability Theory
 - Generalization/Generalizability in Qualitative Research
 - Generalized Additive Models
 - Generalized Estimating Equations
 - Generalized Least Squares
 - Generalized Linear Models
 - Geometric Distribution
 - Gibbs Sampling
 - Gini Coefficient
 - GLIM
 - Going Native
 - Goodness-of-Fit Measures
 - Grade of Membership Models
 - Granger Causality
 - Graphical Modeling
 - Grounded Theory
 - Group Interview
 - Grouped Data
 - Growth Curve Model
 - Guttman Scaling

 - Halo Effect
 - Hawthorne Effect
 - Hazard Rate
 - Hermeneutics
 - Heterogeneity
 - Heteroscedasticity. *See* Heteroskedasticity
 - Heteroskedasticity
 - Heuristic
 - Heuristic Inquiry
 - Hierarchical (Non)Linear Model
 - Hierarchy of Credibility
 - Higher-Order. *See* Order
 - Histogram
 - Historical Methods
 - Holding Constant. *See* Control
 - Homoscedasticity. *See* Homoskedasticity
 - Homoskedasticity
 - Humanism and Humanistic Research
 - Humanistic Coefficient
 - Hypothesis
 - Hypothesis Testing. *See* Significance Testing
 - Hypothetico-Deductive Method

 - Ideal Type
 - Idealism
 - Identification Problem
 - Idiographic/Nomothetic. *See* Nomothetic/Idiographic
 - Impact Assessment
 - Implicit Measures

 - Imputation
 - Independence
 - Independent Variable
 - Independent Variable (in Experimental Research)
 - Independent Variable (in Nonexperimental Research)
 - In-Depth Interview
 - Index
 - Indicator. *See* Index
 - Induction
 - Inequality Measurement
 - Inequality Process
 - Inference. *See* Statistical Inference
 - Inferential Statistics
 - Influential Cases
 - Influential Statistics
 - Informant Interviewing
 - Informed Consent
 - Instrumental Variable
 - Interaction. *See* Interaction Effect;
Statistical Interaction
 - Interaction Effect
 - Intercept
 - Internal Reliability
 - Internal Validity
 - Internet Surveys
 - Interpolation
 - Interpretative Repertoire
 - Interpretive Biography
 - Interpretive Interactionism
 - Interpretivism
 - Interquartile Range
 - Interrater Agreement
 - Interrater Reliability
 - Interrupted Time-Series Design
 - Interval
 - Intervening Variable. *See* Mediating
Variable
 - Intervention Analysis
 - Interview Guide
 - Interview Schedule
 - Interviewer Effects
 - Interviewer Training
 - Interviewing
 - Interviewing in Qualitative Research
 - Intraclass Correlation
 - Intracoder Reliability
 - Investigator Effects
 - In Vivo Coding
 - Isomorph
 - Item Response Theory

- Jackknife Method
- Key Informant
- Kish Grid
- Kolmogorov-Smirnov Test
- Kruskal-Wallis H Test
- Kurtosis
- Laboratory Experiment
- Lag Structure
- Lambda
- Latent Budget Analysis
- Latent Class Analysis
- Latent Markov Model
- Latent Profile Model
- Latent Trait Models
- Latent Variable
- Latin Square
- Law of Averages
- Law of Large Numbers
- Laws in Social Science
- Least Squares
- Least Squares Principle. *See* Regression
- Leaving the Field
- Leslie's Matrix
- Level of Analysis
- Level of Measurement
- Level of Significance
- Levene's Test
- Life Course Research
- Life History Interview. *See* Life Story Interview
- Life History Method
- Life Story Interview
- Life Table
- Likelihood Ratio Test
- Likert Scale
- LIMDEP
- Linear Dependency
- Linear Regression
- Linear Transformation
- Link Function
- LISREL
- Listwise Deletion
- Literature Review
- Lived Experience
- Local Independence
- Local Regression
- Loess. *See* Local Regression
- Logarithm
- Logic Model
- Logical Empiricism. *See* Logical Positivism
- Logical Positivism
- Logistic Regression
- Logit
- Logit Model
- Log-Linear Model
- Longitudinal Research
- Lowess. *See* Local Regression
- Macro
- Mail Questionnaire
- Main Effect
- Management Research
- MANCOVA. *See* Multivariate Analysis of Covariance
- Mann-Whitney U Test
- MANOVA. *See* Multivariate Analysis of Variance
- Marginal Effects
- Marginal Homogeneity
- Marginal Model
- Marginals
- Markov Chain
- Markov Chain Monte Carlo Methods
- Matching
- Matrix
- Matrix Algebra
- Maturation Effect
- Maximum Likelihood Estimation
- MCA. *See* Multiple Classification Analysis
- McNemar Change Test. *See* McNemar's Chi-Square Test
- McNemar's Chi-Square Test
- Mean
- Mean Square Error
- Mean Squares
- Measure. *See* Conceptualization, Operationalization, and Measurement
- Measure of Association
- Measures of Central Tendency
- Median
- Median Test
- Mediating Variable
- Member Validation and Check
- Membership Roles
- Memos, Memoing
- Meta-Analysis
- Meta-Ethnography
- Metaphors
- Method Variance
- Methodological Holism
- Methodological Individualism

-
- Metric Variable
 - Micro
 - Microsimulation
 - Middle-Range Theory
 - Milgram Experiments
 - Missing Data
 - Misspecification
 - Mixed Designs
 - Mixed-Effects Model
 - Mixed-Method Research. *See* Multimethod Research
 - Mixture Model (Finite Mixture Model)
 - MLE. *See* Maximum Likelihood Estimation
 - Mobility Table
 - Mode
 - Model
 - Model I ANOVA
 - Model II ANOVA. *See* Random-Effects Model
 - Model III ANOVA. *See* Mixed-Effects Model
 - Modeling
 - Moderating. *See* Moderating Variable
 - Moderating Variable
 - Moment
 - Moments in Qualitative Research
 - Monotonic
 - Monte Carlo Simulation
 - Mosaic Display
 - Mover-Stayer Models
 - Moving Average
 - Mplus
 - Multicollinearity
 - Multidimensional Scaling (MDS)
 - Multidimensionality
 - Multi-Item Measures
 - Multilevel Analysis
 - Multimethod-Multitrait Research. *See* Multimethod Research
 - Multimethod Research
 - Multinomial Distribution
 - Multinomial Logit
 - Multinomial Probit
 - Multiple Case Study
 - Multiple Classification Analysis (MCA)
 - Multiple Comparisons
 - Multiple Correlation
 - Multiple Correspondence Analysis.
See Correspondence Analysis
 - Multiple Regression Analysis
 - Multiple-Indicator Measures
 - Multiplicative
 - Multistage Sampling
 - Multi-Strategy Research
 - Multivariate
 - Multivariate Analysis
 - Multivariate Analysis of Variance and Covariance (MANOVA and MANCOVA)
 - N (n)
 - N6
 - Narrative Analysis
 - Narrative Interviewing
 - Native Research
 - Natural Experiment
 - Naturalism
 - Naturalistic Inquiry
 - Negative Binomial Distribution
 - Negative Case
 - Nested Design
 - Network Analysis
 - Neural Network
 - Nominal Variable
 - Nomothetic. *See* Nomothetic/Idiographic
 - Nomothetic/Idiographic
 - Nonadditive
 - Nonlinear Dynamics
 - Nonlinearity
 - Nonparametric Random-Effects Model
 - Nonparametric Regression. *See* Local Regression
 - Nonparametric Statistics
 - Nonparticipant Observation
 - Nonprobability Sampling
 - Nonrecursive
 - Nonresponse
 - Nonresponse Bias
 - Nonsampling Error
 - Normal Distribution
 - Normalization
 - NUD*IST
 - Nuisance Parameters
 - Null Hypothesis
 - Numeric Variable. *See* Metric Variable
 - NVivo
 - Objectivism
 - Objectivity
 - Oblique Rotation. *See* Rotations
 - Observation Schedule
 - Observation, Types of
 - Observational Research
 - Observed Frequencies
 - Observer Bias

- Odds
Odds Ratio
Official Statistics
Ogive. *See* Cumulative Frequency Polygon
OLS. *See* Ordinary Least Squares
Omega Squared
Omitted Variable
One-Sided Test. *See* One-Tailed Test
One-Tailed Test
One-Way ANOVA
Online Research Methods
Ontology, Ontological
Open Question. *See* Open-Ended Question
Open-Ended Question
Operational Definition. *See* Conceptualization, Operationalization, and Measurement
Operationism/Operationalism
Optimal Matching
Optimal Scaling
Oral History
Order
Order Effects
Ordinal Interaction
Ordinal Measure
Ordinary Least Squares (OLS)
Organizational Ethnography
Orthogonal Rotation. *See* Rotations
Other, the
Outlier
- p* Value
Pairwise Correlation. *See* Correlation
Pairwise Deletion
Panel
Panel Data Analysis
Paradigm
Parameter
Parameter Estimation
Part Correlation
Partial Correlation
Partial Least Squares Regression
Partial Regression Coefficient
Participant Observation
Participatory Action Research
Participatory Evaluation
Pascal Distribution
Path Analysis
Path Coefficient. *See* Path Analysis
Path Dependence
Path Diagram. *See* Path Analysis
- Pearson's Correlation Coefficient
Pearson's *r*. *See* Pearson's Correlation Coefficient
Percentage Frequency Distribution
Percentile
Period Effects
Periodicity
Permutation Test
Personal Documents. *See* Documents, Types of
Phenomenology
Phi Coefficient. *See* Symmetric Measures
Philosophy of Social Research
Philosophy of Social Science
Photographs in Research
Pie Chart
Piecewise Regression. *See* Spline Regression
Pilot Study
Placebo
Planned Comparisons. *See* Multiple Comparisons
Poetics
Point Estimate
Poisson Distribution
Poisson Regression
Policy-Oriented Research
Politics of Research
Polygon
Polynomial Equation
Polytomous Variable
Pooled Surveys
Population
Population Pyramid
Positivism
Post Hoc Comparison. *See* Multiple Comparisons
Postempiricism
Posterior Distribution
Postmodern Ethnography
Postmodernism
Poststructuralism
Power of a Test
Power Transformations. *See* Polynomial Equation
Pragmatism
Precoding
Predetermined Variable
Prediction
Prediction Equation
Predictor Variable
Pretest
Pretest Sensitization
Priming
Principal Components Analysis
Prior Distribution

-
- Prior Probability
 Prisoner's Dilemma
 Privacy. *See* Ethical Codes; Ethical Principles
 Privacy and Confidentiality
 Probabilistic
 Probabilistic Guttman Scaling
 Probability
 Probability Density Function
 Probability Sampling. *See* Random Sampling
 Probability Value
 Probing
 Probit Analysis
 Projective Techniques
 Proof Procedure
 Propensity Scores
 Proportional Reduction of Error (PRE)
 Protocol
 Proxy Reporting
 Proxy Variable
 Pseudo-*R*-Squared
 Psychoanalytic Methods
 Psychometrics
 Psychophysiological Measures
 Public Opinion Research
 Purposive Sampling
 Pygmalion Effect
- Q Methodology
 Q Sort. *See* Q Methodology
 Quadratic Equation
 Qualitative Content Analysis
 Qualitative Data Management
 Qualitative Evaluation
 Qualitative Meta-Analysis
 Qualitative Research
 Qualitative Variable
 Quantile
 Quantitative and Qualitative Research, Debate About
 Quantitative Research
 Quantitative Variable
 Quartile
 Quasi-Experiment
 Questionnaire
 Queuing Theory
 Quota Sample
- R*
r
 Random Assignment
 Random Digit Dialing
 Random Error
 Random Factor
 Random Number Generator
 Random Number Table
 Random Sampling
 Random Variable
 Random Variation
 Random Walk
 Random-Coefficient Model. *See* Mixed-Effects Model;
 Multilevel Analysis
 Random-Effects Model
 Randomized-Blocks Design
 Randomized Control Trial
 Randomized Response
 Randomness
 Range
 Rank Order
 Rapport
 Rasch Model
 Rate Standardization
 Ratio Scale
 Rationalism
 Raw Data
 RC(M) Model. *See* Association Model
 Reactive Measures. *See* Reactivity
 Reactivity
 Realism
 Reciprocal Relationship
 Recode
 Record-Check Studies
 Recursive
 Reductionism
 Reflexivity
 Regression
 Regression Coefficient
 Regression Diagnostics
 Regression Models for Ordinal Data
 Regression on . . .
 Regression Plane
 Regression Sum of Squares
 Regression Toward the Mean
 Regressor
 Relationship
 Relative Distribution Method
 Relative Frequency
 Relative Frequency Distribution
 Relative Variation (Measure of)
 Relativism
 Reliability
 Reliability and Validity in Qualitative Research

- Reliability Coefficient
Repeated Measures
Repertory Grid Technique
Replication
Replication/Replicability in Qualitative Research
Representation, Crisis of
Representative Sample
Research Design
Research Hypothesis
Research Management
Research Question
Residual
Residual Sum of Squares
Respondent
Respondent Validation. *See* Member Validation and Check
Response Bias
Response Effects
Response Set
Retroduction
Rhetoric
Rho
Robust
Robust Standard Errors
Role Playing
Root Mean Square
Rotated Factor. *See* Factor Analysis
Rotations
Rounding Error
Row Association. *See* Association Model
Row-Column Association. *See* Association Model
R-Squared

Sample
Sampling
Sampling Bias. *See* Sampling Error
Sampling Distribution
Sampling Error
Sampling Fraction
Sampling Frame
Sampling in Qualitative Research
Sampling Variability. *See* Sampling Error
Sampling With (or Without) Replacement
SAS
Saturated Model
Scale
Scaling
Scatterplot
Scheffé's Test
Scree Plot

Secondary Analysis of Qualitative Data
Secondary Analysis of Quantitative Data
Secondary Analysis of Survey Data
Secondary Data
Second-Order. *See* Order
Seemingly Unrelated Regression
Selection Bias
Self-Administered Questionnaire
Self-Completion Questionnaire. *See* Self-Administered Questionnaire
Self-Report Measure
Semantic Differential Scale
Semilogarithmic. *See* Logarithms
Semiotics
Semipartial Correlation
Semistructured Interview
Sensitive Topics, Researching
Sensitizing Concept
Sentence Completion Test
Sequential Analysis
Sequential Sampling
Serial Correlation
Shapiro-Wilk Test
Show Card
Sign Test
Significance Level. *See* Alpha, Significance Level of a Test
Significance Testing
Simple Additive Index. *See* Multi-Item Measures
Simple Correlation (Regression)
Simple Observation
Simple Random Sampling
Simulation
Simultaneous Equations
Skewed
Skewness. *See* Kurtosis
Slope
Smoothing
Snowball Sampling
Social Desirability Bias
Social Relations Model
Sociogram
Sociometry
Soft Systems Analysis
Solomon Four-Group Design
Sorting
Sparse Table
Spatial Regression
Spearman Correlation Coefficient
Spearman-Brown Formula

-
- Specification
Spectral Analysis
Sphericity Assumption
Spline Regression
Split-Half Reliability
Split-Plot Design
S-Plus
Spread
SPSS
Spurious Relationship
Stability Coefficient
Stable Population Model
Standard Deviation
Standard Error
Standard Error of the Estimate
Standard Scores
Standardized Regression Coefficients
Standardized Test
Standardized Variable. *See* z-Score; Standard Scores
Standpoint Epistemology
Stata
Statistic
Statistical Comparison
Statistical Control
Statistical Inference
Statistical Interaction
Statistical Package
Statistical Power
Statistical Significance
Stem-and-Leaf Display
Step Function. *See* Intervention Analysis
Stepwise Regression
Stochastic
Stratified Sample
Strength of Association
Structural Coefficient
Structural Equation Modeling
Structural Linguistics
Structuralism
Structured Interview
Structured Observation
Structured, Focused Comparison
Substantive Significance
Sum of Squared Errors
Sum of Squares
Summated Rating Scale. *See* Likert Scale
Suppression Effect
Survey
Survival Analysis
Symbolic Interactionism
Symmetric Measures
Symmetry
Synchronic
Systematic Error
Systematic Observation. *See* Structured Observation
Systematic Review
Systematic Sampling
Tau. *See* Symmetric Measures
Taxonomy
Tchebechev's Inequality
t-Distribution
Team Research
Telephone Survey
Testimonio
Test-Retest Reliability
Tetrachoric Correlation. *See* Symmetric Measures
Text
Theoretical Sampling
Theoretical Saturation
Theory
Thick Description
Third-Order. *See* Order
Threshold Effect
Thurstone Scaling
Time Diary. *See* Time Measurement
Time Measurement
Time-Series Cross-Section (TSCS) Models
Time-Series Data (Analysis/Design)
Tobit Analysis
Tolerance
Total Survey Design
Transcription, Transcript
Transfer Function. *See* Box-Jenkins Modeling
Transformations
Transition Rate
Treatment
Tree Diagram
Trend Analysis
Triangulation
Trimming
True Score
Truncation. *See* Censoring and Truncation
Trustworthiness Criteria
t-Ratio. *See* *t*-Test
t-Statistic. *See* *t*-Distribution
t-Test
Twenty Statements Test
Two-Sided Test. *See* Two-Tailed Test
Two-Stage Least Squares

- Two-Tailed Test
 Two-Way ANOVA
 Type I Error
 Type II Error
- Unbalanced Designs
 Unbiased
 Uncertainty
 Uniform Association. *See* Association Model
 Unimodal Distribution
 Unit of Analysis
 Unit Root
 Univariate Analysis
 Universe. *See* Population
 Unobserved Heterogeneity
 Unobtrusive Measures
 Unobtrusive Methods
 Unstandardized
 Unstructured Interview
 Utilization-Focused Evaluation
- V.* *See* Symmetric Measures
 Validity
 Variable
 Variable Parameter Models
 Variance. *See* Dispersion
 Variance Components Models
 Variance Inflation Factors
 Variation
 Variation (Coefficient of)
 Variation Ratio
 Varimax Rotation. *See* Rotations
 Vector
- Vector Autoregression (VAR). *See* Granger
 Causality
 Venn Diagram
 Verbal Protocol Analysis
 Verification
Verstehen
 Vignette Technique
 Virtual Ethnography
 Visual Research
 Volunteer Subjects
- Wald-Wolfowitz Test
 Weighted Least Squares
 Weighting
 Welch Test
 White Noise
 Wilcoxon Test
 WinMAX
 Within-Samples Sum of Squares
 Within-Subject Design
 Writing
- X* Variable
 $\bar{X}$
- Y*-Intercept
Y Variable
 Yates' Correction
 Yule's *Q*
- z*-Score
z-Test
 Zero-Order. *See* Order

The Appendix, Bibliography, appears at the end of this volume as well as in Volume 3.

Reader's Guide

ANALYSIS OF VARIANCE

Analysis of Covariance (ANCOVA)
Analysis of Variance (ANOVA)
Main Effect
Model I ANOVA
Model II ANOVA
Model III ANOVA
One-Way ANOVA
Two-Way ANOVA

ASSOCIATION AND CORRELATION

Association
Association Model
Asymmetric Measures
Biserial Correlation
Canonical Correlation Analysis
Correlation
Correspondence Analysis
Intraclass Correlation
Multiple Correlation
Part Correlation
Partial Correlation
Pearson's Correlation
 Coefficient
Semipartial Correlation
Simple Correlation (Regression)
Spearman Correlation Coefficient
Strength of Association
Symmetric Measures

BASIC QUALITATIVE RESEARCH

Autobiography
Life History Method
Life Story Interview
Qualitative Content Analysis
Qualitative Data Management
Qualitative Research

Quantitative and Qualitative Research,
 Debate About
Secondary Analysis of Qualitative Data

BASIC STATISTICS

Alternative Hypothesis
Average
Bar Graph
Bell-Shaped Curve
Bimodal
Case
Causal Modeling
Cell
Covariance
Cumulative Frequency Polygon
Data
Dependent Variable
Dispersion
Exploratory Data Analysis
F Ratio
Frequency Distribution
Histogram
Hypothesis
Independent Variable
Median
Measures of Central Tendency
N(*n*)
Null Hypothesis
Pie Chart
Regression
Standard Deviation
Statistic
t-Test
 $\bar{X}$
Y Variable
z-Test

CAUSAL MODELING

Causality
Dependent Variable

Effects Coefficient
Endogenous Variable
Exogenous Variable
Independent Variable
Path Analysis
Structural Equation Modeling

DISCOURSE/CONVERSATION ANALYSIS

Accounts
Conversation Analysis
Critical Discourse Analysis
Deviant Case Analysis
Discourse Analysis
Foucauldian Discourse Analysis
Interpretative Repertoire
Proof Procedure

ECONOMETRICS

ARIMA
Cointegration
Durbin-Watson Statistic
Econometrics
Fixed Effects Model
Mixed-Effects Model
Panel
Panel Data Analysis
Random-Effects Model
Selection Bias
Serial Correlation (Regression)
Time-Series Cross-Section (TSCS) Models
Time-Series Data (Analysis/Design)
Tobit Analysis

EPISTEMOLOGY

Constructionism, Social
Epistemology
Idealism
Interpretivism
Laws in Social Science
Logical Positivism
Methodological Holism
Naturalism
Objectivism
Positivism

ETHNOGRAPHY

Autoethnography
Case Study
Creative Analytical Practice (CAP) Ethnography

Critical Ethnography
Ethnographic Content Analysis
Ethnographic Realism
Ethnographic Tales
Ethnography
Participant Observation

EVALUATION

Applied Qualitative Research
Applied Research
Evaluation Research
Experiment
Heuristic Inquiry
Impact Assessment
Qualitative Evaluation
Randomized Control Trial

EVENT HISTORY ANALYSIS

Censoring and Truncation
Event History Analysis
Hazard Rate
Survival Analysis
Transition Rate

EXPERIMENTAL DESIGN

Experiment
Experimenter Expectancy Effect
External Validity
Field Experimentation
Hawthorne Effect
Internal Validity
Laboratory Experiment
Milgram Experiments
Quasi-Experiment

FACTOR ANALYSIS AND RELATED TECHNIQUES

Cluster Analysis
Commonality Analysis
Confirmatory Factor Analysis
Correspondence Analysis
Eigenvalue
Exploratory Factor Analysis
Factor Analysis
Oblique Rotation
Principal Components Analysis
Rotated Factor

Rotations
Varimax Rotation

FEMINIST METHODOLOGY

Feminist Ethnography
Feminist Research
Gender Issues
Standpoint Epistemology

GENERALIZED LINEAR MODELS

General Linear Models
Generalized Linear Models
Link Function
Logistic Regression
Logit
Logit Model
Poisson Regression
Probit Analysis

HISTORICAL/COMPARATIVE

Comparative Method
Comparative Research
Documents, Types of
Emic/Etic Distinction
Historical Methods
Oral History

INTERVIEWING IN QUALITATIVE RESEARCH

Biographic Narrative Interpretive
Method (BNIM)
Dependent Interviewing
Informant Interviewing
Interviewing in Qualitative Research
Narrative Interviewing
Semistructured Interview
Unstructured Interview

LATENT VARIABLE MODEL

Confirmatory Factor Analysis
Item Response Theory
Factor Analysis
Latent Budget Analysis
Latent Class Analysis
Latent Markov Model
Latent Profile Model

Latent Trait Models
Latent Variable
Local Independence
Nonparametric Random-Effects Model
Structural Equation Modeling

LIFE HISTORY/BIOGRAPHY

Autobiography
Biographic Narrative Interpretive
Method (BNIM)
Interpretive Biography
Life History Method
Life Story Interview
Narrative Analysis
Psychoanalytic Methods

LOG-LINEAR MODELS (CATEGORICAL DEPENDENT VARIABLES)

Association Model
Categorical Data Analysis
Contingency Table
Expected Frequency
Goodness-of-Fit Measures
Log-Linear Model
Marginal Model
Marginals
Mobility Table
Odds Ratio
Saturated Model
Sparse Table

LONGITUDINAL ANALYSIS

Cohort Analysis
Longitudinal Research
Panel
Period Effects
Time-Series Data (Analysis/Design)

MATHEMATICS AND FORMAL MODELS

Algorithm
Assumptions
Basic Research
Catastrophe Theory
Chaos Theory
Distribution
Fuzzy Set Theory
Game Theory

MEASUREMENT LEVEL

Attribute
Binary
Categorical
Continuous Variable
Dichotomous Variables
Discrete
Interval
Level of Measurement
Metric Variable
Nominal Variable
Ordinal Measure

**MEASUREMENT TESTING
AND CLASSIFICATION**

Conceptualization, Operationalization,
and Measurement
Generalizability Theory
Item Response Theory
Likert Scale
Multiple-Indicator Measures
Summated Rating Scale

MULTILEVEL ANALYSIS

Contextual Effects
Dependent Observations
Fixed Effects Model
Mixed-Effects Model
Multilevel Analysis
Nonparametric Random-Effects
Model
Random-Coefficient Model
Random-Effects Model

MULTIPLE REGRESSION

Adjusted *R*-Squared
Best Linear Unbiased Estimator
Beta
Generalized Least Squares
Heteroskedasticity
Interaction Effect
Misspecification
Multicollinearity
Multiple Regression Analysis
Nonadditive
R-Squared
Regression
Regression Diagnostics

Specification
Standard Error of the Estimate

QUALITATIVE DATA ANALYSIS

Analytic Induction
CAQDAS
Constant Comparison
Grounded Theory
In Vivo Coding
Memos, Memoing
Negative Case
Qualitative Content Analysis

SAMPLING IN QUALITATIVE RESEARCH

Purposive Sampling
Sampling in Qualitative Research
Snowball Sampling
Theoretical Sampling

SAMPLING IN SURVEYS

Multistage Sampling
Quota Sampling
Random Sampling
Representative Sample
Sampling
Sampling Error
Stratified Sampling
Systematic Sampling

SCALING

Attitude Measurement
Bipolar Scale
Dimension
Dual Scaling
Guttman Scaling
Index
Likert Scale
Multidimensional Scaling (MDS)
Optimal Scaling
Scale
Scaling
Semantic Differential Scale
Thurstone Scaling

SIGNIFICANCE TESTING

Alpha, Significance Level of a Test
Confidence Interval

Level of Significance
One-Tailed Test
Power of a Test
Significance Level
Significance Testing
Statistical Power
Statistical Significance
Substantive Significance
Two-Tailed Test

SIMPLE REGRESSION

Coefficient of Determination
Constant
Intercept
Least Squares
Linear Regression
Ordinary Least Squares
Regression on . . .
Regression
Scatterplot
Slope
Y-Intercept

SURVEY DESIGN

Computer-Assisted Personal
Interviewing
Internet Surveys

Interviewing
Mail Questionnaire
Secondary Analysis
of Survey Data
Structured Interview
Survey
Telephone Survey

TIME SERIES

ARIMA
Box-Jenkins Modeling
Cointegration
Detrending
Durbin-Watson Statistic
Error Correction Models
Forecasting
Granger Causality
Interrupted Time-Series Design
Intervention Analysis
Lag Structure
Moving Average
Periodicity
Serial Correlation
Spectral Analysis
Time-Series Cross-Section (TSCS)
Models
Time-Series Data (Analysis/Design)
Trend Analysis

G

GAME THEORY

Game theory extends classical decision theory by focusing on situations in which an actor's result depends not only on his or her own decisions but also on one or more persons' behavior. In this sense, it is a theory of social interaction with a special interest in goal-directed ("rational") behavior.

"Born" by John von Neumann and Oskar Morgenstern (1944), applications of game theory spread from economics to political science, psychology, sociology, and even biology. Key elements of game-theoretic studies are games, players, strategies, payoffs, and maximization. A "game" describes a given situation of social interaction between at least two actors ("players") with all the rules of the game, as well as its possibilities and restrictions. The situation offers the players various alternatives to decide and act ("strategies"). The players' strategies lead to different outcomes along with different utility values ("pay-offs"). Based on these key elements, game-theoretic analysis can show which strategy a player will choose when he or she maximizes his or her payoff. Therefore, game theory is more a mathematical discipline than a psychological theory.

Although goal-directed behavior is a presumption of game theory, theoretical results and empirical evidence sometimes can be bizarre. A famous example is the PRISONER'S DILEMMA (PD). The PD is a simple model of a situation of strategic interdependence with aspects of cooperation and conflict. Table 1 shows the PD that Robert Axelrod (1984) used in his computer tournament. Individually seen, defection always leads

Table 1 Prisoner's Dilemma

		<i>Player 2</i>	
		<i>Cooperation</i>	<i>Defection</i>
<i>Player 1</i>	<i>Cooperation</i>	3,3	0,5
	<i>Defection</i>	5,0	1,1

to a better payoff than cooperation (and is therefore a "dominant strategy"). The dilemma is that although mutual defection causes a worse result than mutual cooperation, a player has no interest in a one-sided change of his or her strategy. With this characterization, the PD can be a fruitful model of various social situations such as employment relations, an arms race, or free trade. It also illustrates that game-theoretical analysis is most interesting in situations with mixed motives of cooperation and conflict incentives.

Looking at classical methods of data collection, most empirical game-theoretical studies use EXPERIMENTS, COMPUTER SIMULATIONS, and—in a broader sense—CONTENT ANALYSIS. In a content analysis, given documents such as statistics or historical data are analyzed to build a game-theoretic model (e.g., an explanation of political strategic moves at the beginning of World War I). Computer simulations can match a variety of different strategies to simulate the effect of repeated games (iterations), changed parameters, or evolution aspects. Experiments are the best way of testing (game) theoretical findings. The study by Anatol Rapoport and Albert Chammah (1965) about the PD exemplifies the connection of theoretical, experimental, and simulation aspects. For a general introduction, see Gintis (2000).

—Bernhard Prosch

REFERENCES

- Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.
- Gintis, H. (2000). *Game theory evolving*. Princeton, NJ: Princeton University Press.
- Rapoport, A., & Chammah, A. (1965). *Prisoner's dilemma*. Ann Arbor: Michigan University Press.
- von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behaviour*. Princeton, NJ: Princeton University Press.

GAMMA

Similar to Kendall's tau coefficient (and its variations such as Kendall's τ_b coefficient or Stuart's τ_c coefficient), described in SYMMETRIC MEASURES, Goodman and Kruskal's gamma coefficient is one of the NONPARAMETRIC STATISTICS that describes the relationship between two ordered or rank-order variables X and Y . When there are no tied ranks, all the above coefficients yield the same results. When many tied ranks occur in either variable, the gamma coefficient is more useful and appropriate than tau correlations and will always be greater than Kendall's coefficients.

Given two rank-order variables, X and Y , in a population, the parameter gamma (γ) is defined as the difference between the probability that agreement exists in the order for a pair (X_i, Y_j) and the probability that disagreement exists for a pair (X_i, Y_j) in their ordering, given that there are no ties in the data (Siegel & Castellan, 1988). In other words,

$$\gamma = \frac{p(\text{agreement between } X_i \text{ and } Y_i) - p(\text{disagreement between } X_i \text{ and } Y_i)}{1 - p(X_i \text{ and } Y_i \text{ are tied})}$$

$$= \frac{p(\text{agreement between } X_i \text{ and } Y_i) - p(\text{disagreement between } X_i \text{ and } Y_i)}{p(\text{agreement between } X_i \text{ and } Y_i) + p(\text{disagreement between } X_i \text{ and } Y_i)}.$$

The gamma coefficient of a population can be estimated by a sample using $\hat{\gamma} = \frac{P-Q}{P+Q}$, where P is the number of concordance or agreement, and Q is the number of discordance or disagreement in the sample. The numbers of concordance and discordance are operationalized as follows. In any pair of observations— (X_i, Y_i) for case i and (X_j, Y_j) for case j —concordance occurs when $(X_i - X_j)(Y_i - Y_j) > 0$, which suggests either $Y_i < Y_j$ when $X_i < X_j$ or

Table 1 Ranks of 10 Students on Creative Writing by Two Judges

	Judge X	Judge Y	P	Q
Mr. Lubinsky	1	5	4	5
Ms. Firth	2	4	4	4
Ms. Aron	3	3	4	3
Mr. Millar	4	2	4	2
Ms. Savident	5	1	3	0
Mr. DeLuca	5	1	3	0
Ms. Pavarotti	5	6	3	0
Ms. Lin	6	7	1	0
Mr. Stewart	6	7	1	0
Ms. Barton	7	8	0	0
Σ			27	14

NOTE: P = the number of concordance or agreement; Q = the number of discordance or disagreement.

$Y_i > Y_j$ when $X_i > X_j$. Likewise, discordance exists if $(X_i - X_j)(Y_i - Y_j) < 0$, which suggests either $Y_i < Y_j$ when $X_i > X_j$ or $Y_i > Y_j$ when $X_i < X_j$. The gamma coefficient can range from -1 when $P = 0$ to 1 when $Q = 0$.

An example of the use of this statistic is provided by a situation in which two judges (X and Y) rank 10 students according to their creative writing, with 1 as the lowest rank and 10 as the highest rank. To calculate the gamma coefficient, we first arrange Judge X's rankings in ascending order, as shown in the second column of Table 1. The corresponding rankings from Judge Y are listed in the third column of Table 1. Next, we determine the number of concordance and discordance for each student, starting with the first. Mr. Lubinsky is compared to 9 other students (i.e., Lubinsky vs. Firth, Lubinsky vs. Aron, . . . , and Lubinsky vs. Barton). The number of concordance (i.e., the number of students ranked above 5 by Judge Y) is four, and the number of discordance (i.e., the number of students ranked below 5 by Judge Y) is five. The second student, Ms. Firth, is then compared to the 8 remaining students (i.e., Firth vs. Aron, Firth vs. Millar, . . . , and Firth vs. Barton). The number of concordance (i.e., the number of the remaining students ranked above 4 by Judge Y) is four, and the number of discordance (i.e., the number of the remaining students ranked below 4 by Judge Y) is four.

When there are tied ranks assigned by Judge X, no comparison is made among those cases. For instance, the fifth student, Ms. Savident, is only compared to the remaining students, excluding Mr. DeLuca and Ms. Pavarotti (i.e., Savident vs. Lin, Savident vs.

Stewart, and Savident vs. Barton). The number of concordance (i.e., the number of the remaining students ranked above 1 by Judge Y) is three, and the number of discordance (i.e., the number of the remaining students ranked below 1 by Judge Y) is 0. Following the same procedure, both P and Q for the remaining students can be obtained.

The null hypothesis test for Goodman-Kruskal's gamma can be examined by

$$z = \frac{\hat{\gamma}}{\sqrt{\frac{n(1-\hat{\gamma}^2)}{P+Q}}}$$

when the sample size is large. Note that the upper limit of the standard error is

$$\sqrt{\frac{n(1-\hat{\gamma}^2)}{P+Q}}.$$

Therefore, the z -statistic obtained from the above formula is a conservative estimate.

—Peter Y. Chen and Autumn D. Krauss

REFERENCE

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

GAUSS-MARKOV THEOREM. See BLUE; OLS

GEISTESWISSENSCHAFTEN

This is a term that came into common use in the 19th century, referring to the humanities and social sciences. In the context of methodology, it also implies a particular approach to work in these fields, one that emphasizes the role of VERSTEHEN.

There were isolated uses of *Geisteswissenschaften*, or its variants, before 1849, but widespread employment of the term seems to have stemmed from Schiel's translation into German of John Stuart Mill's *System of Logic* (Mill, 1974), in which the term *Geisteswissenschaften* was used to stand for Mill's "moral sciences." Mill regarded psychology

as the key moral science, and its method was to be modeled on the natural sciences (*Naturwissenschaften*). However, most of the German historians and philosophers who adopted the term *Geisteswissenschaften* took a very different view. Some earlier German writers, notably Goethe and Hegel, had sought to develop a notion of scientific understanding that encompassed not just natural science but also the humanities and philosophy, thereby opposing empiricist and materialist views of scientific method and their extension to the study of human social life. However, by the middle of the 19th century, attempts at such a single, comprehensive view of scholarship had been largely discredited in many quarters. Instead, influential philosophers argued that the human social world had to be studied in a quite *different* way from the manner in which natural scientists investigate the physical world, although one that had equal empirical rigor. As a result, it is common to find *Geisteswissenschaften* translated back into English as "the human studies" or "the human sciences," signaling distance from the positivist view of social research marked by labels such as *behavioral science*.

One of the most influential accounts of the distinctiveness of the *Geisteswissenschaften* is found in the work of Wilhelm Dilthey (see Ermarth, 1978). He argued that whereas the natural sciences properly approach the study of phenomena from the outside, by studying their external characteristics and behavior, we are able to understand human beings and their products—whether artifacts, beliefs, ways of life, or whatever—from the inside because we share a common human nature with those we study. Another influential 19th-century account of the distinctiveness of the human sciences was put forward by the neo-Kantians Windelband and Rickert. For them, the distinction between the natural and social sciences was not primarily about subject matter; indeed, the neo-Kantians were keen to emphasize that there is only one reality (see Hammersley, 1989, pp. 28–33). Rather, the main difference was between a generalizing and an individualizing form of science. Max Weber drew on both the Diltheyan and neo-Kantian traditions in developing his *verstehende* sociology, in which explanation (*Erklären*) and understanding (*Verstehen*) go together, so that any sociological account is required to be adequate at the level of both cause and meaning (Turner & Factor, 1981).

—Martyn Hammersley

REFERENCES

- Ermarth, M. (1978). *Wilhelm Dilthey: The critique of historical reason*. Chicago: University of Chicago Press.
- Hammersley, M. (1989). *The dilemma of qualitative method: Herbert Blumer and the Chicago school of sociology*. London: Routledge.
- Mill, J. S. (1974). *Collected works of John Stuart Mill: Vol. 8. A system of logic* (Books 4–6). Toronto: University of Toronto Press.
- Turner, S. P., & Factor, R. A. (1981). Objective possibility and adequate causality in Weber's methodological writings. *Sociological Review*, 29(1), 5–28.

GENDER ISSUES

Issues relating to gender in social research arose, in the 1970s, out of concerns about the relative invisibility of women and a perceived lack of attention to their lives. Feminists argued that there had been insufficient recognition of how differences in the experiences of women and men, together with changes in the relationships between them, could have significant implications for the ways in which social phenomena are conceptualized and investigated. Areas researched tended to focus on the public spheres of life, those existing outside the private sphere of the home, such as the workplace and school. It was assumed either that the experiences of women could simply be inferred from those of men or that men had the normal or customary kinds of lives, with those of women being treated as unusual or deviant.

GENDER AND RESEARCH

One way of dealing with the gender issue is to treat women and men as separate and to measure the extent of any differences between them, for example, in health and illness or educational attainment. Such an approach, although useful, has been criticized as inadequate on its own because it fails to acknowledge differences in the quality and nature of experiences. Another approach, characteristic of the early stages of gender awareness, was to add studies of women onto those of men. However, although valuable work was produced in areas such as employment and the media, for instance, this was similarly criticized for insufficiently challenging the basic assumptions and concepts that informed such research. It was argued that these were still based on male perceptions and

interests, with women being tagged on rather as an afterthought. This led to three developments. First, it was necessary to research additional issues that were particularly pertinent to women. New themes were introduced, ranging from housework, motherhood, and childbirth to sexuality, pornography, and male violence. Second, some of the perspectives and concepts at the very heart of the social science agenda had to be revised or modified. This gave rise to important debates, such as the relative meaning and significance of social class for women and men and whether the accepted distinction between work and leisure was based on gendered assumptions. The third development involved issues to do with the actual methods and procedures of social research. There was heated discussion among feminists, with some suggesting that qualitative, rather than quantitative, research styles were more attuned to rendering the previously hidden aspects of women's lives visible. There has also been much debate as to whether it is either desirable or possible to have a distinct feminist methodology (Ramazanoglu, 2002).

DEFINING AND OPERATIONALIZING GENDER

The most usual definition of gender distinguishes it from biological sex. Whereas sex is taken to denote the anatomical and physiological differences that define biological femaleness and maleness, gender refers to the social construction of femininity and masculinity (Connell, 2002). It is not regarded as a direct product of biological sex but is seen as being differently created and interpreted by particular cultures and in particular periods of historical time. However, a number of issues arise from this usage. To begin with, initial studies of gender relations focused mainly on women and girls to compensate for their previous exclusion. It is only recently that the gendered nature of men, of masculinities, has become the focus of attention. Furthermore, the commonsense usage of the term *gender* has been contested. It has been suggested that there are not, in practice, two mutually exclusive gender categories, as attempts to classify international athletes have demonstrated (Ramazanoglu, 2002). Intersexuality and transsexuality disrupt rigid gender boundaries, indicating that body differences may be important after all. Definitions based on dichotomy also overlook patterns of difference among both women and men—for instance, between violent and nonviolent masculinities. Another concern is with the need to relate gender

to other forms of difference—for instance, those of ethnicity, sexual orientation, disability, class, and age. Such categories and signifiers mean that gender is always mediated in terms of experience.

There is also the postmodern argument that gender has no pre-given and essential existence. It only exists to the extent that it continues to be performed. Gender is rendered diverse, fluctuating, and volatile rather than a stable and homogeneous identity. Such a stance, however, sweeps away the grounds for focusing on gender issues in the first place. Despite all the above problems, the researcher on gender has to have a more specific place to start. This has led some commentators to talk about research on “gendered lives” (Ramazanoglu, 2002, p. 5) and the need to focus on gender “relations” (Connell, 2002, p. 9) rather than accept gender as a given. This makes the understanding of what constitutes gender a part of any particular empirical investigation. It also makes it a form of patterning of social relations within which individuals and groups act. The central argument is that gender cannot be known in general or in advance of research.

EFFECTS ON THE RESEARCH PROCESS

It has also been claimed that gender can influence the research process, whatever the topic being studied and whether this is from a feminist or nonfeminist perspective. For example, differing gendered perceptions of the social world may influence why a particular project is chosen, the ways in which it is framed, and the questions that are then asked. The gender of the researcher, together with other characteristics, such as age and ethnicity, can affect the nature of the interaction with research participants and the quality of the information obtained. Gender may also be of significance in the analysis and interpretation of data, with men and women “reading” material in differing ways, based on their differing experiences and worldview. There is debate in the literature as to how far these gender effects may be minimized. Views range from those who see them to be inconsequential for research to those who regard them as possible sources of bias. For the latter, it is necessary to be reflexive about the way gender might intrude into research practice and also to make commentary on it available so that this can be part of any critical evaluation process.

—Mary Maynard

REFERENCES

- Connell, R. W. (2002). *Gender*. Cambridge, UK: Polity.
 Ramazanoglu, C. (with Holland, J.). (2002). *Feminist methodology*. London: Sage.

GENERAL LINEAR MODELS

The general linear model provides a basic framework that underlies many of the statistical analyses that social scientists encounter. It is the foundation for many univariate, multivariate, and repeated-measures procedures, including the *t*-TEST, (MULTIVARIATE) ANALYSIS OF VARIANCE, (MULTIVARIATE) ANALYSIS OF COVARIANCE, MULTIPLE REGRESSION ANALYSIS, FACTOR ANALYSIS, CANONICAL CORRELATION ANALYSIS, CLUSTER ANALYSIS, DISCRIMINANT ANALYSIS, MULTIDIMENSIONAL SCALING, REPEATED-MEASURES DESIGN, and other analyses. It can be succinctly expressed in a regression-type formula, for example, in terms of the expected value of the random dependent variable Y , such as income as a function of a number of independent variables X_k (for $k = 1$ to K), such as age, education, and social class:

$$E(Y) = f(X_1, \dots, X_K) \\ = \beta_0 + \beta_1 X_1 + \dots + \beta_K X_K. \quad (1)$$

All members of the general linear model can be given in the form of (1). For example, a typical analysis of variance model is formed when all the explanatory variables are categorical factors and hence can be coded as DUMMY VARIABLES in X_k . Parameters in the general linear model can be appropriately obtained by ORDINARY LEAST SQUARES estimation. According to the Gauss-Markov theorem, when all the ASSUMPTIONS of the general linear model are satisfied, such estimation is optimal in that the least squares ESTIMATORS are BEST LINEAR UNBIASED ESTIMATORS of the population parameters of β_k .

A basic distinction between the general linear model and the GENERALIZED LINEAR MODEL lies in the response variable of the model. The former requires a continuous variable, whereas the latter does not. Another distinction is the distribution of the response variable, which is expected to be normal in the general linear model, but it can be one of several nonnormal (e.g., BINOMIAL, MULTINOMIAL, and POISSON DISTRIBUTION) distributions. Yet another distinction is how the response variable is related to the explanatory

variables. In the generalized linear model, the relation between the dependent and the independent variables is specified by the LINK FUNCTION, an example of which is the identity link of (1), which is the only link used in the general linear model. On the other hand, the link function of the generalized linear model, which is formally expressed by the inverse function in (1) or $f^1[E(Y)]$, can take on nonlinear links such as the probit, the logit, and the log, among others:

$$\begin{aligned} E(Y) &= f(X_1, \dots, X_K) \\ &= f(\beta_0 + \beta_1 X_1 + \dots + \beta_K X_K). \end{aligned} \quad (2)$$

Here the identity link is just one of the many possibilities, and the general linear model (1) can be viewed as a special case of the generalized linear model of (2).

Another extension of the general linear model is the GENERALIZED ADDITIVE MODEL, which also extends the generalized linear model by allowing for nonparametric functions. The generalized additive model can be expressed similarly in terms of the expected value of Y :

$$\begin{aligned} E(Y) &= f(X_1, \dots, X_K) \\ &= f[s_0 + s_1(X_1) + \dots + s_K(X_K)], \end{aligned} \quad (3)$$

where $s_K(X_K)$ are smoothing functions that are estimated nonparametrically. In (3), not only the $f^1[E(Y)]$ link function can take on various nonlinear forms, but each of the X_K variables is fitted using a smoother such as the B-spline.

Another type of generalization that can be applied to the models of (1) and (2) captures the spirit of the HIERARCHICAL LINEAR MODEL or the models in MULTI-LEVEL ANALYSIS. So far, we have ignored the subscript i representing individual cases. If the subscript i is used for the first-level units and j for the second-level, or higher, units, then the Y_{ij} and X_{ij} variables represent observations from two levels; moreover, the coefficients β_{jk} now are random and can be distinguishable among the second-level units.

Some popular statistical software packages such as SAS and SPSS implement the general linear model in a flexible procedure for data analysis. Both SAS and STATA have procedures for fitting generalized linear and generalized additive models as well as some form of hierarchical (non)linear models.

—Tim Futing Liao

GENERALIZABILITY. See EXTERNAL VALIDITY

GENERALIZABILITY THEORY

Generalizability theory provides an extensive conceptual framework and set of statistical machinery for quantifying and explaining the consistencies and inconsistencies in observed scores for objects of measurement. To an extent, the theory can be viewed as an extension of classical test theory through the application of certain ANALYSIS OF VARIANCE (ANOVA) procedures. Classical theory postulates that an observed score can be decomposed into a TRUE SCORE and a single undifferentiated RANDOM ERROR term. By contrast, generalizability theory substitutes the notion of *universe score* for true score and liberalizes classical theory by employing ANOVA methods that allow an investigator to untangle the multiple sources of error that contribute to the undifferentiated error in classical theory.

Perhaps the most important and unique feature of generalizability theory is its conceptual framework, which focuses on certain types of studies and universes. A *generalizability* (G) study involves the collection of a sample of data from a *universe of admissible observations* that consists of *facets* defined by an investigator. Typically, the principal result of a G study is a set of estimated random-effects variance components for the universe of admissible observations. A D (*decision*) study provides estimates of universe score variance, error variances, certain indexes that are like RELIABILITY COEFFICIENTS, and other statistics for a measurement procedure associated with an investigator-specified *universe of generalization*.

In univariate generalizability theory, there is only one universe (of generalization) score for each object of measurement. In multivariate generalizability theory, each object of measurement has multiple universe scores. Univariate generalizability theory typically employs random-effects models, whereas MIXED-EFFECTS MODELS are more closely associated with multivariate generalizability theory.

HISTORICAL DEVELOPMENT

The defining treatment of generalizability theory is the 1972 book by Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam titled *The Dependability of Behavioral Measurements*. A recent extended treatment of the theory is the 2001 book by Robert L. Brennan titled *Generalizability Theory*. The essential features of univariate generalizability theory were largely completed with technical reports in 1960–1961 that were revised into three journal articles, each with a different first author (Cronbach, Gleser, and Rajaratnam). Multivariate generalizability theory was developed during the ensuing decade. Research between 1920 and 1955 by Ronald A. Fisher, Cyril Burt, Cyril J. Hoyt, Robert L. Ebel, and E. F. Lindquist, among others, influenced the development of generalizability theory.

EXAMPLE

Suppose an investigator, Smith, wants to construct a measurement procedure for evaluating writing proficiency. First, Smith might identify or characterize essay prompts of interest, as well as potential raters. Doing so specifies the facets in the universe of admissible observations. Assume these facets are viewed as infinite and that, in theory, any person p (i.e., object of measurement) might respond to any essay prompt t , which in turn might be evaluated by any rater r . If so, a likely data collection design might be $p \times t \times r$, where “ $\times$ ” is read as “crossed with.” The associated linear model is

$$X_{ptr} = \mu + \nu_p + \nu_t + \nu_r + \nu_{pt} + \nu_{pr} + \nu_{tr} + \nu_{ptr},$$

where the ν are uncorrelated score effects. This model leads to

$$\sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r) + \sigma^2(pt) + \sigma^2(pr) + \sigma^2(tr) + \sigma^2(ptr), \quad (1)$$

which is a decomposition of the total observed score variance into seven *variance components* that are usually estimated using the expected MEAN SQUARES in a random-effects ANOVA.

Suppose the following estimated variance components are obtained from Smith’s G study based on n_t essay prompts and n_r raters:

$$\begin{aligned} \hat{\sigma}^2(p) &= 0.25, & \hat{\sigma}^2(t) &= 0.06, & \hat{\sigma}^2(r) &= 0.02, \\ \hat{\sigma}^2(pt) &= 0.15, & \hat{\sigma}^2(pr) &= 0.04, & \hat{\sigma}^2(tr) &= 0.00, \\ & & \text{and } \hat{\sigma}^2(ptr) &= 0.12. \end{aligned}$$

Suppose also that Smith wants to generalize persons’ observed mean scores based on n'_t essay prompts and n'_r raters to these persons’ scores for a universe of generalization that involves an infinite number of prompts and raters. This is a verbal description of a D study with a $p \times T \times R$ random-effects design. It is much like the $p \times t \times r$ design for Smith’s G study, but the sample sizes for the D study need not be the same as the sample sizes for the G study, and the $p \times T \times R$ design focuses on *mean* scores for persons. The expected score for a person over the facets in the universe of generalization is called the person’s universe score.

If $n'_t = 3$ and $n'_r = 2$ for Smith’s measurement procedure, then the estimated random-effects variance components for the D study are

$$\begin{aligned} \hat{\sigma}^2(p) &= 0.25, & \hat{\sigma}^2(T) &= \frac{\hat{\sigma}^2(t)}{n'_t} = 0.02, \\ \hat{\sigma}^2(R) &= \frac{\hat{\sigma}^2(r)}{n'_r} = 0.01, & \hat{\sigma}^2(pT) &= \frac{\hat{\sigma}^2(pt)}{n'_t} = 0.05, \\ \hat{\sigma}^2(pR) &= \frac{\hat{\sigma}^2(pr)}{n'_r} = 0.02, & \hat{\sigma}^2(TR) &= \frac{\hat{\sigma}^2(tr)}{n'_t n'_r} = 0.00, \\ & & \text{and } \hat{\sigma}^2(pTR) &= \frac{\hat{\sigma}^2(ptr)}{n'_t n'_r} = 0.02, \end{aligned}$$

with $\hat{\sigma}^2(p) = .25$ as the estimated universe score variance. The other variance components contribute to different types of error variance.

Absolute error, Δ_p , is simply the difference between a person’s observed mean score and the person’s universe score. For Smith’s design and universe, the variance of these errors is

$$\begin{aligned} \sigma^2(\Delta) &= \sigma^2(T) + \sigma^2(R) + \sigma^2(pT) + \sigma^2(pR) \\ &\quad + \sigma^2(TR) + \sigma^2(pTR), \end{aligned}$$

which gives $\hat{\sigma}^2(\Delta) = 0.02 + 0.01 + 0.05 + 0.02 + 0.00 + 0.02 = 0.12$. Its square root is $\hat{\sigma}(\Delta) = 0.35$, which is interpretable as an estimate of the “absolute” standard error of measurement for a randomly selected person.

Relative error, δ_p , is defined as the difference between a person’s observed deviation score and his or

her universe deviation score. For Smith's design and universe,

$$\sigma^2(\delta) = \sigma^2(pT) + \sigma^2(pR) + \sigma^2(pTR),$$

which gives $\hat{\sigma}^2(\delta) = 0.05 + 0.02 + 0.02 = 0.09$. Its square root is $\hat{\sigma}(\delta) = 0.30$, which is interpretable as an estimate of the "relative" standard error of measurement for a randomly selected person.

Two types of reliability-like coefficients are widely used in generalizability theory. One coefficient is called a *generalizability coefficient*, which is defined as

$$E\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}.$$

It is the analog of a reliability coefficient in classical theory. The other coefficient is called an *index of dependability*, which is defined as

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)}.$$

Because $\sigma^2(\delta) < \sigma^2(\Delta)$, it necessarily follows that $E\rho^2 > \Phi$. For example, using Smith's data, $E\rho^2 = 0.25/(0.25 + 0.09) = 0.74$, and $\hat{\Phi} = 0.25/(0.25 + 0.12) = 0.68$.

—Robert L. Brennan

REFERENCES

- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

GENERALIZATION/ GENERALIZABILITY IN QUALITATIVE RESEARCH

The issue of whether one should or can generalize from the findings of qualitative (interpretive/ethnographic) research lies at the heart of the approach, although it is discussed less than one would expect. Its relative absence from the methodological literature is likely attributable to it being seen by many

qualitative researchers as a preoccupation of those who take a NOMOTHETIC, specifically quantitative, approach to research. For many qualitative researchers, it is a nonissue because the role of qualitative research is to interpret the meanings of agents within particular social contexts, rather than measuring, explaining, or predicting.

Those who deny the possibility of generalization argue that individual consciousnesses are free to attach different meanings to the same actions or circumstances. Conversely, different actions can arise out of similarly expressed meanings. An "inherent indeterminateness in the lifeworld" (Denzin, 1983, p. 133) is said to follow, which leads to too much variability to allow the possibility of generalization from a specific situation to others. However, a major problem in holding this line is that it is actually very hard not to generalize, in that reports of a particular social setting, events, or behavior will usually claim some typicality (or, conversely, that the reports are in some way untypical of some prior observed regularity).

Those who deny the possibility or desirability of generalization would be correct in their assertions if it was claimed that qualitative research could produce total generalizations, where situation S^{1-n} is identical to S in every detail, or even statistical generalizations, where the probability of situation S occurring more widely can be estimated from instances of s . The first kind of generalization is rarely possible outside of physics or chemistry, and the second (in social science) is the province of survey research. However, it seems reasonable to claim that qualitative research might produce *moderatum* generalizations in which aspects of S can be seen to be instances of a broader recognizable set of features (Williams, 2000). These depend on levels of cultural consistency in the social environment, those things that are the basis of inductive reasoning in everyday life, such as rules, customs, and shared social constructions of the physical environment.

Moderatum generalizations can arise from CASE STUDY research. Case studies can take many different forms—for example, the identification of the existence of a general social principle, the analysis of a social situation, or the extended case study in which people or events are studied over time (Mitchell, 1983, pp. 193–194). Case studies do not depend on statistical generalization from sample to population, as in survey research, but on logical inference from prior theorizing. Theoretical generalization therefore does not aim

to say anything about populations but instead makes claims about the existence of phenomena proposed by a theory. This reasoning is common in anthropology, and theoretical corroboration is said to be increased by further instances of a phenomenon in repeated case studies.

Some minimal generalization seems necessary if qualitative research is to have any value in the policy process. If all qualitative research can give us are many interpretations, each with the same epistemological status, then how can we evaluate, for example, social programs or economic policy? Nevertheless, one should be cautious when making even moderate generalizations from qualitative data (as analogously as one should in claims of meaning in survey data). Even if findings confirm a prior theoretical proposition, it cannot be known how typical they are with any certainty.

—Malcolm Williams

REFERENCES

- Denzin, N. (1983). Interpretive interactionism. In G. Morgan (Ed.), *Beyond method: Strategies for social research* (pp. 128–142). Beverly Hills, CA: Sage.
- Mitchell, J. (1983). Case and situation analysis. *Sociological Review*, 31(2), 186–211.
- Williams, M. (2000). Interpretivism and generalisation. *Sociology*, 34(2), 209–224.

GENERALIZED ADDITIVE MODELS

Additive models recast the LINEAR REGRESSION model $y_i = \alpha + \sum_{j=1}^k X_{ij}\beta_j + \varepsilon_i$ by modeling y as an additive combination of nonparametric functions of X : $y_i = \alpha + \sum_{j=1}^k m_j(X_{ij}) + \varepsilon_i$, where the ε_i are zero mean, independent, and identically distributed stochastic disturbances. *Generalized* additive models (GAMs) extend additive models in precisely the same way the generalized linear models (GLMs) extend linear models so as to accommodate qualitative dependent variables (e.g., binary outcomes, count data, etc.). The additive model for a continuous outcome is a special case of a GAM, just as ORDINARY LEAST SQUARES regression with normal disturbances is a special type of GLM.

GAMs are especially useful in exploratory work or when analysts have weak priors as to the functional

form relating predictors X to an outcome y . In these cases, assuming that the conditional mean function for y , $E(y|X)$ is linear in X is somewhat arbitrary: Although linearity is simple, easy to interpret, and parsimonious, linear functional forms risk smoothing over what might be substantively interesting nonlinearities in $E(y|X)$. GAMs are also useful when prediction motivates the data analysis. By fitting possibly quite nonlinear $m(\cdot)$ functions, the resulting model will respond to localized features of the data, resulting in smaller prediction errors than those from a linear model.

GAMs can be distinguished from other nonparametric modeling strategies (for a recent comparison, see Hastie, Tibshirani, & Friedman, 2001). With additive GLMs at one extreme and NEURAL NETWORKS at the other, GAMs lie closer to the former than the latter, generally retaining the additivity (and hence separability) of linear regression. GAMs are also generally closer to linear regression models than other techniques such as projection pursuit, alternating conditional expectations, or tree-based methods. GAMs can also be distinguished from parametric approaches to handling nonlinearities, such as simply logging variables, modeling with polynomials of the predictors, or the Box-Cox transformation; the $m(\cdot)$ functions that GAMs estimate are nonparametric functions that are generally more flexible than parametric approaches. Note also that linear models nest as a special case of a GAM, meaning that the null hypothesis of a linear fit can be tested against a nonparametric, nonlinear alternative; that is, the null hypothesis is that $m(X) = X\beta$, or, in other words, a 1 degree-of-freedom $m(\cdot)$ versus a higher degree-of-freedom alternative. In practice, a mixture of parametric and nonparametric $m(\cdot)$ functions may be fit, say, when researchers have strong beliefs that a particular predictor has a linear effect on y but wish to fit a flexible $m(\cdot)$ function tapping the effect of some other predictor.

GAMs use *linear smoothers* as their $m(\cdot)$ functions, predictor by predictor. It is possible to fit an $m(\cdot)$ function over more than a single predictor, but to simplify the discussion, I consider the case of fitting nonparametric functions predictor by predictor: Given n observations on a dependent variable $\mathbf{y} = (y_1, \dots, y_n)'$ and a predictor $\mathbf{x} = (x_1, \dots, x_n)'$, a linear smoother estimates $\hat{y}$ at target point x_0 as a weighted average (i.e., a linear function) of the y observations accompanying the x observations in some neighborhood of the target point x_0 . That is, $\hat{y} = \hat{m}(x_0) = \sum_{i=1}^n S(x_0, x_i; \lambda)y_i$,

where $S(\cdot)$ is a weighting function, parameterized by a SMOOTHING parameter λ , that effectively governs the width of the neighborhood around the target point. The neighborhood width or *bandwidth* governs how much weight observations $i = 1, \dots, n$ contribute to forming $\hat{y}$ at target point x_0 : As the name implies, local fitting assigns zero weight to data points not in the specified neighborhood of the target point x_0 , and it is the smoothing parameter λ that determines “how local” is “local.”

To motivate these ideas, one should consider LOESS, a popular linear smoother by Cleveland (1979). Given a target point x_0 , loess yields $\hat{y}|x_0 = \hat{m}(x_0)$ by fitting a *locally weighted* regression, using the following procedure:

1. Identify the q nearest neighbors of x_0 (i.e., the q value of x closest to x_0). This set is denoted $N(x_0)$. The parameter q is the equivalent of the smoothing parameter λ , defined above, and is usually preset by the analyst.
2. Calculate $\Delta(x_0) = \max N(x_0)|x_0 - x_i|$, the distance of the nearest neighbor most distant from x_0 .
3. Calculate weights w_i for each point in $N(x_0)$ using the following tri-cube weight function:

$$W \left[\frac{|x_0 - x_i|}{\Delta(x_0)} \right],$$

where

$$W(z) = \begin{cases} (1 - z^3)^3 & \text{for } 0 \leq z < 1; \\ 0 & \text{otherwise.} \end{cases}$$

Note that $W(x_0) = 1$ and that the weights decline smoothly to zero over the chosen set of nearest neighbors, such that $W(x_i) = 0 \forall x_i \notin N(x_0)$. The use of the tri-cube weight function is somewhat arbitrary; any weight function that has smooth contact with 0 on the boundary of $N(x_0)$ will produce a smooth fitted function.

4. Regress y on x and an intercept (for local linear fitting) using WEIGHTED LEAST SQUARES, with weights W defined in the previous step. Quadratic or higher order polynomial local regression can be implemented (Fan & Gijbels, 1996, chap. 5) or even mixtures of low-order polynomials. Local constant fitting produces a simple moving average.

5. The predicted value from the local weighted regression is the smoothed value $\hat{y}|x_0 = \hat{m}(x_0)$.

Loess is just one of a number of nonparametric smoothers but demonstrates the key idea; run a regression of y and x in the neighborhood around a target point x_0 (smoothers differ as to how to weight observations in the neighborhood of x_0 —that is, the choice of the *kernel* of the nonparametric regression—and in the degree of the local regression), and then report the fitted value of the local regression at x_0 as the estimate of the nonparametric function. Other popular smoothers for nonparametric regression include SPLINE functions, loess, and various other smoothing functions. The spline and the loess, among a variety of smoothers (including the kernel), are typically available software implementing GAMs.

Implementing the linear smoother over a grid of target points traces out a function, the smoothed fit of y against x . In particular, letting the linear smoother produce $\hat{y}_i$ for each observation (i.e., letting $x_0 = x_i, i = 1, \dots, n$) produces $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, where $\mathbf{S}$ is an n -by- n smoother matrix, formed by stacking n vectors of weights $S(x_i, x_i; \lambda)$, each of length n . Linear smoothers are thus linear in precisely the same way that the predicted values from a least squares regression are a linear function of $\mathbf{y}$: that is, $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the well-known “hat matrix.” The smoother matrix $\mathbf{S}$ has many of the same properties as $\mathbf{H}$. In fact, $\mathbf{H}$ is a special case of a linear smoother, and least squares linear regression is the equivalent of an infinitely smooth smoother (formally, “smoothness” is measured by the inverse of the second derivative of the fitted values with respect to the predictors; for a linear model, this second derivative is zero). At the other extreme is a smoother that simply interpolates the data (the equivalent of running a regression with a dummy variable for every observation), and $\mathbf{S} = \mathbf{I}_n$ (an n -by- n identity matrix), with all off-diagonal elements zero (i.e., extreme local fitting). As the bandwidth increases, an increasing number of the off-diagonal elements of $\mathbf{S}$ are nonzero, as data points further away from any target point x_i have greater influence in forming $\hat{y}_i = \hat{m}(x_i)$.

The inverse relationship between the amount of smoothing and the extent to which $\mathbf{S}$ is diagonal means that the degrees of freedom consumed by a linear smoother are well approximated by the trace (sum of the diagonal elements) of $\mathbf{S}$ (Hastie & Tibshirani, 1990, Appendix 2). Because all smoothers fit an intercept term, $\text{tr}(\mathbf{S}) - 1$ is an estimate of the *equivalent degrees of freedom* consumed by the nonparametric component of the linear smoother. This approximation

provides a critical ingredient in inference, letting the smoother's residual error variance σ^2 be estimated as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n [y_i - \hat{m}(x_i)]^2}{n - \text{tr}(\mathbf{S})},$$

and, in turn, the variance-covariance matrix of the smoothed values is estimated with $\text{var}[\hat{\mathbf{y}}] = \mathbf{S}\mathbf{S}'\hat{\sigma}^2$, with pointwise standard errors for the fitted smooth obtained by taking the square root of the diagonal of this matrix. In short, the more local the fitting, the less smoothing and the more degrees of freedom are consumed, with less bias in the $\hat{\mathbf{y}}$ but with the price of higher variances.

GAMs extend linear smoothers to multiple predictors via an iterative fitting ALGORITHM known as backfitting, described in detail in Hastie and Tibshirani (1990). The basic idea is to cycle over the predictors in the model, $j = 1, \dots, k$, smoothing the residual quantity y minus the intercept and the partial effects of the $k - 1$ other predictors against x_j until convergence. The extension to GLM-type models with qualitative dependent variables is also reasonably straightforward; backfitting nests neatly inside the iteratively reweighted least squares algorithm for GLMs developed by Nelder and Wedderburn (1972).

A key question when working with GAMs is whether a simpler, more parsimonious fit will suffice relative to the fitted model. This is critical in helping to guard against overfitting (fitting a more flamboyant function than the data support) and in finding an appropriate setting for the respective smoothing parameters. Simpler, "less local" models (including a linear fit) will nest inside an alternative, "more local" fit, suggesting that likelihood ratio tests can be used to test the null hypothesis that the restrictions embodied in the simpler model are correct. These tests apply only approximately in the nonparametric setting implied by GAMs (for one thing, the degrees of freedom consumed by the GAM are only an estimate) but, in large samples, appear to work well. There are a number of proposals for automatic selection of smoothing parameters in the statistical literature (e.g., Hastie & Tibshirani, 1990, pp. 50–52), and this is an active area of research.

As a practical matter, social scientists should be aware that implementing GAMs entails forgoing a simple single-parameter assessment of the marginal effect of x on y produced by a linear regression or a GLM. In adopting a nonparametric approach, the best way (and perhaps only way) to appreciate the marginal effect of x on y is via a graph of the fitted smooth

of y against x . Beck and Jackman (1998) provide numerous examples from political science, supplementing traditional tables of regression estimates with graphs of nonparametric/nonlinear marginal effects and confidence intervals around the fitted values.

GAMs are the subject of a book by Hastie and Tibshirani (1990), although the literature on nonparametric smoothing (at the heart of any GAM) is immense. Loader (1999) provides an excellent review of local fitting methods; Wahba (1990) reviews the use of splines in nonparametric regression. GAMs are implemented in S-PLUS (for a detailed review, see Hastie, 1992). Wood (2001) provides an implementation in R with automatic selection of the smoothing parameters via generalized cross-validation and smoothing via spline functions.

—Simon Jackman

REFERENCES

- Beck, N., & Jackman, S. (1998). Beyond linearity by default: Generalized additive models. *American Journal of Political Science*, 42, 596–627.
- Cleveland, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.
- Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman & Hall.
- Hastie, T. J. (1992). Generalized additive models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S* (pp. 249–307). Pacific Grove, CA: Wadsworth Brooks/Cole.
- Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman & Hall.
- Hastie, T. J., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.
- Loader, C. (1999). *Local regression and likelihood*. New York: Springer.
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135, 370–384.
- Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.
- Wood, S. (2001). mgcv: GAMs and generalized ridge regression for R. *R News*, 1, 20–25.

GENERALIZED ESTIMATING EQUATIONS

The method of generalized estimating equations (GEE) is an extension of GENERALIZED LINEAR MODELS

to repeated-measures (or, in fact, any correlated) data. In the social sciences, this class of MODELS is most valuable for PANEL and TIME-SERIES CROSS-SECTION MODEL data, although it can be used to model spatial data as well.

Consider a generalized linear model for repeated-measures data, where $i = (1, \dots, N)$ indexes units and $t = (1, \dots, T)$ indexes time points, such that $E(Y_{it}) \equiv \mu_{it} = f(\mathbf{X}_{it}\beta)$ and $\text{Var}(Y_{it}) = g(\mu_{it})/\varphi$, where $\mathbf{X}$ is an $(NT \times k)$ vector of covariates, $f(\cdot)$ is the inverse of the LINK FUNCTION, and φ is a scale parameter. The GEE estimate of β is the solution to the set of k differential “score” equations:

$$\mathbf{U}_k(\beta) = \sum_{i=1}^N \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mu_i) = \mathbf{0}, \quad (1)$$

where

$$\mathbf{V}_i = \frac{\mathbf{A}_i^{1/2} \mathbf{R}_i(\alpha) \mathbf{A}_i^{1/2}}{\varphi}.$$

$\mathbf{A}_i$ is a $t \times t$ diagonal matrix with $g(\mu_{it})$ as its t th diagonal element, and $\mathbf{R}_i(\alpha)$ is a $t \times t$ “working” correlation matrix of the conditional within-subject correlation among the values of Y_{it} . The form of the working correlation matrix is determined by the researcher; however, Liang and Zeger (1986) show that GEE estimates of β and its robust variance-covariance matrix are consistent even in the presence of MISSPECIFICATION of $\mathbf{R}_i(\alpha)$. In practice, researchers typically specify $\mathbf{R}_i(\alpha)$ according to their substantive beliefs about the nature of the within-subject correlations.

GEEs are a robust and attractive alternative to other models for panel and time-series cross-sectional data. In particular, GEE methods offer the possibility of estimating models in which the intrasubject CORRELATION is itself of substantive interest. Such “GEE2” methods (e.g., Prentice & Zhao, 1991) incorporate a second, separate set of estimating equations for the pairwise intrasubject correlations. In addition, GEEs offer the ability to model continuous, dichotomous, ordinal, and event-count data using a wide range of distributions. Finally, GEEs are estimable using a wide range of commonly used software packages, including SAS, S-PLUS, and STATA.

—Christopher Zorn

REFERENCES

- Hardin, J. W., & Hilbe, J. M. (2003). *Generalized estimating equations*. London: Chapman & Hall.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.
- Prentice, R. L., & Zhao, L. P. (1991). Estimating equations for parameters in mean and covariances of multivariate discrete and continuous responses. *Biometrics*, 47, 825–839.
- Zorn, C. J. W. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, 45, 470–490.

GENERALIZED LEAST SQUARES

Consider the LINEAR REGRESSION model $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, where $\mathbf{y}$ is a n -by-1 vector of observations on a dependent variable, $\mathbf{X}$ is a n -by- k matrix of independent variables of full column rank, β is a k -by-1 vector of parameters to be estimated, and $\mathbf{u}$ is a n -by-1 vector of disturbances. Via the GAUSS-MARKOV theorem, if

Assumption 1 (A1): $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$ (i.e., the disturbances have conditional mean zero), and

Assumption 2 (A2): $E(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \sigma^2\mathbf{\Omega}$, where $\mathbf{\Omega} = \mathbf{I}_n$, an n -by- n identity matrix (i.e., conditional on the $\mathbf{X}$, the disturbances are independent and identically distributed or “iid” with conditional variance σ^2),

then the ORDINARY LEAST SQUARES (OLS) estimator $\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ with variance-covariance matrix $V(\hat{\beta}_{OLS})\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is (a) the BEST LINEAR UNBIASED ESTIMATOR (BLUE) of β , in the sense of having smallest sampling variability in the class of linear unbiased estimators, and (b) a consistent estimator of β (i.e., as $n \rightarrow \infty$, $\Pr[|\hat{\beta}_{OLS} - \beta| < \varepsilon] = 1$, for any $\varepsilon > 0$, or $\text{plim } \hat{\beta}_{OLS} = \beta$).

If A2 fails to hold (i.e., $\mathbf{\Omega}$ is a positive definite matrix but not equal to $\mathbf{I}_n$), then $\hat{\beta}_{OLS}$ remains unbiased, but no longer “best,” and remains consistent. Relying on $\hat{\beta}_{OLS}$ when A2 does not hold risks faulty inferences; without A2, $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ is a biased and inconsistent estimator of $V(\hat{\beta}_{OLS})$, meaning that the estimated STANDARD ERRORS for $\hat{\beta}_{OLS}$ are wrong, risking invalid inferences and hypothesis tests. A2 often fails to hold in practice; for example, (a) pooling across disparate units often generates disturbances with different conditional

variances (HETEROSKEDASTICITY), and (b) analysis of TIME-SERIES data often generates disturbances that are not conditionally independent (serially correlated disturbances).

When A2 does not hold, it may be possible to implement a *generalized least squares* (GLS) estimator that is BLUE (at least asymptotically). For instance, if the researcher knows the exact form of the departure from A2 (i.e., the researcher knows Ω), then the GLS estimator $\hat{\beta}_{\text{GLS}} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}$ is BLUE, with variance-covariance matrix $\sigma^2(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}$. Note that when A2 holds, $\Omega = \mathbf{I}_n$, and $\hat{\beta}_{\text{GLS}} = \hat{\beta}_{\text{OLS}}$ (i.e., OLS is a special case of the more general estimator).

Typically, researchers do not possess exact knowledge of Ω , meaning that $\hat{\beta}_{\text{GLS}}$ is nonoperational, and an *estimated or feasible* generalized least squares (FGLS) estimator is used. FGLS estimators are often implemented in multiple steps: (a) an OLS analysis to yield estimated residuals $\hat{\mathbf{u}}$; (b) an analysis of the $\hat{\mathbf{u}}$ to form an estimate of Ω , denoted $\hat{\Omega}$; and (c) computation of the FGLS estimator $\hat{\beta}_{\text{FGLS}} = (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{y}$. The third step is often performed by noting that $\hat{\Omega}$ can be decomposed as $\hat{\Omega} = \mathbf{P}^{-1}(\mathbf{P}')^{-1}$, and $\hat{\beta}_{\text{FGLS}}$ is obtained by running the WEIGHTED LEAST SQUARES (WLS) regression of $\mathbf{y}^* = \mathbf{P}\mathbf{y}$ on $\mathbf{X}^* = \mathbf{P}\mathbf{X}$; that is, $\hat{\beta}_{\text{FGLS}} = (\mathbf{X}^*\mathbf{X}^*)^{-1}\mathbf{X}^*\mathbf{y}^*$.

The properties of FGLS estimators vary depending on the form of Ω (i.e., the nature of the departure from the conditional iid assumption in A2) and the quality of $\hat{\Omega}$, and so they cannot be neatly summarized. Finite sample properties of the FGLS estimator are often derived case by case via MONTE CARLO experiments; in fact, it is possible to find cases in which $\hat{\beta}_{\text{OLS}}$ is more efficient than $\hat{\beta}_{\text{FGLS}}$, say, when the violation of A2 is mild (e.g., Chipman, 1979; Rao & Griliches, 1969). ASYMPTOTIC results are more plentiful and usually rely on showing that $\hat{\beta}_{\text{FGLS}}$ and $\hat{\beta}_{\text{GLS}}$ are asymptotically equivalent, so that $\hat{\beta}_{\text{FGLS}}$ is a consistent and asymptotically efficient estimator of β (e.g., Amemiya, 1985, pp. 186–222; Judge, Griffiths, Hill, & Lee, 1980, pp. 117–118).

In social science settings, FGLS is most commonly encountered in dealing with simple forms of residual autocorrelation and heteroskedasticity. For instance, the popular Cochrane-Orcutt (Cochrane & Orcutt, 1949) and Prais-Winsten (Prais & Winsten, 1954) procedures for AR(1) disturbances yield FGLS estimators. The use of FGLS to deal with

heteroskedasticity appears in many econometrics texts: Judge et al. (1980, pp. 128–145) and Amemiya (1985, pp. 198–207) provide rigorous treatments of FGLS estimators of β for commonly encountered forms of heteroskedasticity. Groupwise heteroskedasticity arises from pooling across disparate units, yielding disturbances that are conditionally iid *within* groups of observations; this model is actually a special case of Zellner's (1962) SEEMINGLY UNRELATED REGRESSION (SUR) model. FGLS applied to Zellner's SUR model results in a cross-equation weighting scheme that is also used as the third stage of three-stage least squares estimators (Zellner & Theil, 1962).

FGLS is also used to estimate models for PANEL data with unit- and/or period-specific HETEROGENEITY not captured by the independent variables—for example, $y_{it} = \mathbf{x}_{it}'\beta + v_i + \eta_t + \varepsilon_{it}$, where v_i and η_t are error components (or random effects) specific to units i and time periods t , respectively. The composite error term $u_{it} = v_i + \eta_t + \varepsilon_{it}$ is generally not iid, but its variance-covariance matrix Ω can be estimated with the sum of the estimated variances of the error components; provided that $E(u_{it}|\mathbf{x}_{it}) = 0$ and $E(\mathbf{X}'\Omega^{-1}\mathbf{X})$ is of full rank, then FGLS provides a consistent estimator of β (see Wooldridge, 2002, chap. 10).

Finally, FGLS estimators can often be obtained as MAXIMUM LIKELIHOOD ESTIMATES (MLEs), generating consistent estimates of β in “one shot.” Although the iid assumption (A2) often greatly simplifies deriving and computing MLEs, the more common (and hence more simple) departures from A2 are often tapped via a small number of auxiliary parameters that determine the content and structure of Ω (e.g., consider the relatively simple form of Ω implied by first-order residual autocorrelation, groupwise heteroskedasticity, or simple error components structures). Thus, the parameters of substantive interest (β) and parameters tapping the assumed error process can be estimated simultaneously via MLE, yielding estimates that are asymptotically equivalent to FGLS estimates.

—Simon Jackman

REFERENCES

- Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MA: Harvard University Press.
- Chipman, J. S. (1979). Efficiency of least squares estimation of linear trend when residuals are autocorrelated. *Econometrica*, 47, 115–128.

- Cochrane, D., & Orcutt, G. H. (1949). Application of least squares relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, 44, 32–61.
- Judge, G. G., Griffiths, W. E., Hill, R. C., & Lee, T.-C. (1980). *The theory and practice of econometrics*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley.
- Prais, S. J., & Winsten, C. B. (1954). *Trend estimators and serial correlation*. Chicago: Cowles Commission.
- Rao, P., & Griliches, Z. (1969). Small-sample properties of several two-stage regression methods in the context of auto-correlated errors. *Journal of the American Statistical Association*, 64, 253–272.
- Wooldridge, J. M. (2002). *Econometric analysis of cross-section and panel data*. Cambridge: MIT Press.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57, 348–368.
- Zellner, A., & Theil, H. (1962). Three-stage least squares: Simultaneous estimation of simultaneous equations. *Econometrica*, 30, 54–78.

GENERALIZED LINEAR MODELS

Generalized linear models (GLMs) expand the basic structure of the well-known linear model to accommodate nonnormal and noninterval measured outcome variables in a single unified theoretical form. It is common in the social sciences to encounter outcome variables that do not fit the standard assumptions of the linear model, and as a result, many distinct forms of regression have been developed: POISSON REGRESSION for counts as outcomes, LOGIT and PROBIT ANALYSIS for dichotomous explanations, exponential models for DURATIONS, gamma models for TRUNCATED data, and more. Although these forms are commonly used and well understood, it is not widely known among practitioners that nearly all of these particularistic regression techniques are special forms of the *generalized linear model*.

With the generalized linear model, explanatory variables are treated in exactly the usual fashion by creating a linear systematic component, $\eta = \mathbf{X}\beta$, which defines the right-hand side of the statistical model. Because the expectation of the outcome variable, $\mu = -\mathbf{y}$, is no longer from an interval-measured form, standard central limit and asymptotic theory no longer apply in the same way, and a “LINK” FUNCTION is required to define the relationship between the linear systematic component of the data and the mean of the outcome

variable, $\mu = g(\mathbf{X}\beta)$. If the specified link function is simply the identity function, $g(\mathbf{X}\beta) = \mathbf{X}\beta$, then the generalized linear model reduces to the linear model, hence its name.

The generalized linear model is based on well-developed theory, starting with Nelder and Wedderburn (1972) and McCullagh and Nelder (1989), which states that any parametric form for the outcome variable that can be recharacterized (algebraically) into the *exponential family form* leads to a link function that connects the mean function of this parametric form to the linear systematic component. This exponential family form is simply a standardized means of expressing various probability mass functions (PMFs, for discrete forms) and PROBABILITY DENSITY FUNCTIONS (PDFs, for continuous forms).

The value of the GLM approach is that a seemingly disparate set of nonlinear regression models can be integrated into a single framework in which the processes of specification searching, numerical estimation, residuals analysis, goodness-of-fit analysis, and reporting are all unified. Thus, researchers and practitioners can learn a singularly general procedure that fits a wide class of data types and can also be developed to include even broader classes of assumptions than those described here.

THE EXPONENTIAL FAMILY FORM

The key to understanding the generalized linear model is knowing how common probability density functions for continuous data forms and probability mass functions for discrete data forms can be expressed in exponential family form—a form that readily produces link functions and moment statistics. The exponential family form originates with Fisher (1934, pp. 288–296), but it was not shown until later that members of the exponential family have fixed dimension-sufficient statistics and produce unique MAXIMUM LIKELIHOOD ESTIMATES with strong consistency and asymptotic normality (Gill, 2000, p. 22).

Consider the one-parameter conditional PDF or PMF for the random variable Z of the form $f(z|\zeta)$. It is classified as being an exponential family form if it can be expressed as

$$f(z|\zeta) = \exp[\underbrace{t(z)u(\zeta)}_{\text{interaction component}} + \underbrace{\log(r(z)) + (\log s(\zeta))}_{\text{additive component}}],$$

(1)

given that r and t are real-valued functions of z that do not depend on ζ , and s and u are real-valued functions of ζ that do not depend on z , with $r(z) > 0, s(\zeta) > 0 \forall z, \zeta$. The key feature is the distinctiveness of the subfunctions within the exponent. The piece labeled *interaction component* must have $t(z)$, strictly a function of z , multiplying $u(\zeta)$, strictly a function of ζ . Furthermore, the *additive component* must have similarly distinct, but now summed, subfunctions with regard to z and ζ .

The canonical form of the exponential family expression is a simplification of (1) that reveals greater structure and gives a more concise summary of the data. This one-to-one transformation works as follows. Make the transformations $y = t(z)$ and $\theta = u(\zeta)$, if necessary (i.e., they are not already in canonical form), and now express (1) as

$$f(y|\theta) = \exp[y\theta - b(\theta) + c(y)]. \quad (2)$$

The subfunctions in (2) have specific identifiers: y is the canonical form for the data (and typically a sufficient statistic as well), θ is the natural canonical form for the unknown parameter, and $b(\theta)$ is often called the “cumulant function” or “normalizing constant.” The function $c(y)$ is usually not important in the estimation process. However, $b(\theta)$ plays a key role in both determining the mean and variance functions as well as the link function. It should also be noted that the canonical form is invariant to random sampling, meaning that it retains its functionality for iid (independent, identically distributed) data:

$$f(y|\theta) = \exp\left[\sum_{i=1}^n y_i \theta - nb(\theta) + \sum_{i=1}^n c(y_i)\right],$$

and under the extension to multiple parameters through a k -length parameters vector $\boldsymbol{\theta}$:

$$f(y|\boldsymbol{\theta}) = \exp\left[\sum_{j=1}^k (y\theta_j - b(\theta_j)) + c(y)\right].$$

As an example of this theory, we can rewrite the familiar binomial PMF,

$$f(y|n, p) = \binom{n}{y} p^y (1-p)^{n-y},$$

in exponential family form:

$$f(y|n, p) = \exp\left[\underbrace{y \log\left(\frac{p}{1-p}\right)}_{y\theta} - \underbrace{(-n \log(1-p))}_{b(\theta)} + \underbrace{\log\left(\binom{n}{y}\right)}_{c(y)}\right]$$

Here the subfunctions are labeled to correspond to the canonical form previously identified.

THE EXPONENTIAL FAMILY FORM AND MAXIMUM LIKELIHOOD

To obtain estimates of the unknown k -dimensional $\boldsymbol{\theta}$ vector, given an observed matrix of data values $f(\boldsymbol{\theta}|\mathbf{X})$, we can employ the standard technique of maximizing the likelihood function with regard to coefficient values to find the “most likely” values of the $\boldsymbol{\theta}$ vector (Fisher, 1925, pp. 707–709). Asymptotic theory assures us that for sufficiently large samples, the likelihood surface is unimodal in dimensions for exponential family forms, so the maximum likelihood estimate (MLE) process is equivalent to finding the k -dimensional mode.

Regarding $f(\mathbf{X}|\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ for given observed (now fixed) data $\mathbf{X}$, then $L(\boldsymbol{\theta}|\mathbf{X}) = f(\mathbf{X}|\boldsymbol{\theta})$ is called a likelihood function. The MLE, $\hat{\boldsymbol{\theta}}$, has the characteristic that $L(\hat{\boldsymbol{\theta}}|\mathbf{X}) \geq L(\boldsymbol{\theta}|\mathbf{X}) \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the admissible range of $\boldsymbol{\theta}$ given by the parametric form assumptions.

It is more convenient to work with the natural log of the likelihood function, and this does not change any of the resulting parameter estimates because the likelihood function and the log-likelihood function have identical modal points. Using (2) with a scale parameter added, the likelihood function in canonical exponential family notation is

$$\begin{aligned} \ell(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}) &= \log(f(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi})) \\ &= \log\left(\exp\left[\frac{\mathbf{y}\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\boldsymbol{\psi})} + c(\mathbf{y}, \boldsymbol{\psi})\right]\right) \\ &= \frac{\mathbf{y}\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\boldsymbol{\psi})} + c(\mathbf{y}, \boldsymbol{\psi}). \end{aligned} \quad (3)$$

Because all of the terms are expressed in the exponent of the exponential family form, removing this exponent through the log of the likelihood gives a very compact form. The score function is the first derivative of the log-likelihood function with respect to the parameters of interest. For the time being, the scale parameter ψ is treated as a *nuisance parameter* (not of primary interest), and we estimate a scalar. The resulting score function, denoted as $\dot{\ell}(\theta|\psi, \mathbf{y})$, is

$$\begin{aligned}\dot{\ell}(\theta, \psi|\mathbf{y}) &= \frac{\partial}{\partial \theta} \ell(\theta|\psi, \mathbf{y}) \\ &= \frac{\partial}{\partial \theta} \left[\frac{\mathbf{y}\theta - b(\theta)}{a(\psi)} + \mathbf{c}(\mathbf{y}, \psi) \right] \\ &= \frac{\mathbf{y} - \frac{\partial}{\partial \theta} b(\theta)}{a(\psi)}.\end{aligned}\quad (4)$$

The mechanics of the maximum likelihood process involve setting $\dot{\ell}(\theta|\psi, \mathbf{y})$ equal to zero and then solving for the parameter of interest, giving the MLE: $\hat{\theta}$. Furthermore, the *likelihood principle* states that once the data are observed, all of the available evidence from the data for estimating $\hat{\theta}$ is contained in the calculated likelihood function, $\ell(\theta|\psi, \mathbf{y})$.

The value of identifying $b(\theta)$ the function lies in its direct link to the mean and variance functions. The expected value calculation of (2) with respect to the data (Y) in the notation of (3) is

$$\begin{aligned}E_Y \left[\frac{y - \frac{\partial}{\partial \theta} b(\theta)}{a(\psi)} \right] &= 0, \\ \int_Y y f(y) dy - \int_Y \frac{\partial b(\theta)}{\partial \theta} f(y) dy &= 0, \\ \underbrace{\int_Y y f(y) dy}_{E(Y)} - \frac{\partial b(\theta)}{\partial \theta} \underbrace{\int_Y f(y) dy}_1 &= 0, \\ \Rightarrow E[Y] &= \frac{\partial}{\partial \theta} b(\theta),\end{aligned}$$

where the second to last step requires general regularity conditions with regard to the bounds of integration, and all exponential family distributions meet this requirement (Gill, 2000, p. 23). This means that (2) gives the mean of the associated exponential family of distributions: $\mu = b'(\theta)$. It can also be shown (Gill, 2000, p. 26) that the second derivative of $b(\theta)$ is the variance function. Then, multiplying this by the scale

parameter (if appropriate) and substituting back in for the canonical parameter produces the standard variance calculation. These calculations are summarized for the binomial PMF:

Table 1 Binomial Mean and Variance Functions

$b(\theta)$	$n \log(1 + \exp(\theta))$
$E[Y] = \frac{\partial}{\partial \theta} b(\theta) _{\theta=u(\zeta)}$	np
$\frac{\partial^2}{\partial \theta^2} b(\theta)$	$n \exp(\theta)(1 + \exp(\theta))^{-2}$
$\text{Var}[Y] = a(\psi) \frac{\partial^2}{\partial \theta^2} b(\theta) _{\theta=u(\zeta)}$	$np(1 - p)$

THE GENERALIZED LINEAR MODEL THEORY

Start with the standard linear model meeting the following Gauss-Markov conditions:

$$\begin{aligned}\mathbf{V} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ (n \times 1) & \quad (n \times p)(p \times 1) \quad (n \times 1), \\ E(\mathbf{V}) &= \boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}. \quad (5) \\ (n \times 1) & \quad (n \times 1) \quad (n \times p)(p \times 1)\end{aligned}$$

The right-hand sides of the two equations in (5) contain the following: $\mathbf{X}$, the matrix of observed data values; $\mathbf{X}\boldsymbol{\beta}$, the “linear structure vector”; and $\boldsymbol{\varepsilon}$, the error terms. The left-hand side contains $E(\mathbf{V}) = \boldsymbol{\theta}$, the vector of means (i.e., the systematic component). The variable, $\mathbf{V}$, is distributed iid normal with mean θ and constant variance σ^2 . Now suppose we generalize this with a new “linear predictor” based on the mean of the outcome variable, which is no longer required to be normally distributed or even continuous:

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}.$$

(n × 1) (n × 1) (n × p)(p × 1)

It is important here that $g(\cdot)$ be an invertible, *smooth* function of the mean vector $\boldsymbol{\mu}$.

The effect of the explanatory variables is now expressed in the model only through the link from the linear structure, $\mathbf{X}\boldsymbol{\beta}$, to the linear predictor, $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, controlled by the form of the link function, $g(\cdot)$. This link function connects the linear predictor to the *mean* of the outcome variable and not directly to the expression of the outcome variable itself, so the outcome variable can now take on a variety of nonnormal forms. The link function connects the stochastic component, which describes some response variable from a wide variety of forms to all of the standard normal theory

supporting the linear systematic component through the mean function:

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

$$g^{-1}(g(\boldsymbol{\mu})) = g^{-1}\boldsymbol{\eta} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\mu} = E(\mathbf{Y}).$$

Furthermore, this linkage is provided by the form of the mean function from the exponential family subfunction: $b(\theta)$.

For example, with the binomial PMF, we saw that $b(\theta) = n \log(1 + \exp(\theta))$, and $\theta = \log(\frac{p}{1-p})$. Reexpressing this in link function notation is just $\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \log(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}})$, or we could give the inverse link, $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = \frac{\exp(\boldsymbol{\eta})}{1 + \exp(\boldsymbol{\eta})}$.

The generalization of the linear model as described now has the following components:

1. *Stochastic component*: $\mathbf{Y}$ is the random or stochastic component that remains distributed iid according to a specific exponential family distribution with mean $\boldsymbol{\mu}$.

2. *Systematic component*: $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ is the systematic component with an associated Gauss-Markov normal basis.

3. *Link function*: The stochastic component and the systematic component are linked by a function of $\boldsymbol{\eta}$, which is taken from the inverse of the canonical link, $b(\theta)$.

4. *Residuals*: Although the residuals can be expressed in the same manner as in the standard linear model—observed outcome variable value minus predicted outcome variable value—a more useful quantity is the deviance residual described as follows.

ESTIMATING GENERALIZED LINEAR MODELS

Unlike the standard linear model, estimating generalized linear models is not done with a closed-form analytical expression [i.e., $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$]. Instead, the maximum likelihood estimate of the unknown parameter is found with a weighted numerical process that repeatedly increases the likelihood with improved weights on each cycle: *iterative weighted least squares* (IWLS). This process, first proposed by Nelder and Wedderburn (1972, pp. 372–374) and first implemented in the GLIM software package, works for any GLM based on an exponential family form (and for some others).

Currently, all professional-level statistic computing implementations now employ IWLS to numerically find maximum likelihood estimates for generalized linear models.

The overall strategy is to apply Newton-Raphson with Fisher scoring to the normal equations, which is equivalent to iteratively applied, weighted *least squares* (and hence easy). Define the current (or starting) point of the linear predictor by

$$\hat{\boldsymbol{\eta}}_0 = \mathbf{X}'\boldsymbol{\beta}_0$$

$(n \times 1) \quad (n \times p)(p \times 1)$

with fitted value $\hat{\boldsymbol{\mu}}_0$ from $g^{-1}(\hat{\boldsymbol{\eta}}_0)$. Form the “adjusted dependent variable” according to

$$\mathbf{z}_0 = \hat{\boldsymbol{\eta}}_0 + \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}} \Big|_{\hat{\boldsymbol{\mu}}_0} \right) (\mathbf{y} - \hat{\boldsymbol{\mu}}_0),$$

$(n \times 1) \quad (n \times 1) \quad \text{diag}(n \times n) \quad (n \times 1)$

which is a linearized form of the link function applied to the data. As an example of this derivative function, the binomial form looks like

$$\boldsymbol{\eta} = \log\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right) \Rightarrow \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}} = (\boldsymbol{\mu})^{-1}(1-\boldsymbol{\mu})^{-1}.$$

Now form the *quadratic weight matrix*, which is the variance of $\mathbf{z}$:

$$\boldsymbol{\omega}_0^{-1} = \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}} \Big|_{\hat{\boldsymbol{\mu}}_0} \right)^2 \nu(\boldsymbol{\mu}) \Big|_{\hat{\boldsymbol{\mu}}_0},$$

$\text{diag}(n \times n) \quad \text{diag}(n \times n)$

where $\nu(\boldsymbol{\mu})$ is the following variance function: $\frac{\partial}{\partial \boldsymbol{\theta}} b'(\boldsymbol{\theta}) = b''(\boldsymbol{\theta})$. Also, note that this process is necessarily iterative because both $\mathbf{z}$ and $\boldsymbol{\omega}$ depend on the current fitted value, $\boldsymbol{\mu}_0$. The general scheme can now be summarized in the three steps:

1. Construct $\mathbf{z}$, $\boldsymbol{\omega}$. Regress $\mathbf{z}$ on the covariates with weights to get a new interim estimate:

$$\hat{\boldsymbol{\beta}}_1 = \left(\begin{matrix} \mathbf{X}' & \boldsymbol{\omega}_0 & \mathbf{X} \end{matrix} \right)^{-1} \begin{matrix} \mathbf{X}' & \boldsymbol{\omega}_0 & \mathbf{z}_0 \\ (p \times n) & (n \times n) & (n \times 1) \end{matrix}.$$

2. Use the coefficient vector estimate to update the linear predictor:

$$\hat{\boldsymbol{\eta}}_1 = \mathbf{X}'\hat{\boldsymbol{\beta}}_1.$$

3. Iterate the following:

$$\begin{aligned} \mathbf{z}_1 \boldsymbol{\omega}_1 &\Rightarrow \hat{\boldsymbol{\beta}}_2, \hat{\eta}_2 \\ \mathbf{z}_2 \boldsymbol{\omega}_2 &\Rightarrow \hat{\boldsymbol{\beta}}_3, \hat{\eta}_3 \\ \mathbf{z}_3 \boldsymbol{\omega}_3 &\Rightarrow \hat{\boldsymbol{\beta}}_4, \hat{\eta}_4 \\ &\vdots \end{aligned}$$

Under very general conditions, satisfied by the exponential family of distributions, the iterative weighted least squares procedure finds the mode of the likelihood function, thus producing the maximum likelihood estimate of the unknown coefficient vector, $\hat{\boldsymbol{\beta}}$. Furthermore, the matrix produced by $\hat{\sigma}^2(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}$ converges in probability to the variance matrix of $\hat{\boldsymbol{\beta}}$ as desired.

RESIDUALS AND DEVIANCES

One significant advantage of the generalized linear model is the freedom it provides from the standard Gauss-Markov assumption that the residuals have mean zero and constant variance. Unfortunately, this freedom comes with the price of interpreting more complex stochastic structures. The *response* residual vector, $\mathbf{R}_{\text{response}} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, calculated for linear models can be updated to include the GLM link function, $\mathbf{R}_{\text{response}} = \mathbf{Y} - g^{-1}(\mathbf{X}\boldsymbol{\beta})$, but this does not then provide the nice distribution theory we get from the standard linear model.

A more useful but related idea for generalized linear models is the *deviance function*. This is built in a similar fashion to the likelihood ratio statistic, comparing the log likelihood from a proposed model specification to the maximum log likelihood possible through the saturated model (n data points and n specified parameters, using the same data and link function). The resulting difference is multiplied by 2 and called the summed deviance (Nelder & Wedderburn, 1972, pp. 374–376). The goodness-of-fit intuition is derived from the idea that this sum constitutes the summed contrast of individual likelihood contributions with the native data contributions to the saturated model. The point here is to compare the log likelihood for the proposed model,

$$\ell(\hat{\boldsymbol{\theta}}, \psi | \mathbf{y}) = \sum_{i=1}^n \frac{y_i \hat{\boldsymbol{\theta}} - b(\hat{\boldsymbol{\theta}})}{a(\psi)} + c(\mathbf{y}, \psi),$$

to the same log-likelihood function with identical data and the same link function, except that it is now with n coefficients for the n data points (i.e., the saturated model log-likelihood function):

$$\ell(\tilde{\boldsymbol{\theta}}, \psi | \mathbf{y}) = \sum_{i=1}^n \frac{y_i \tilde{\boldsymbol{\theta}} - b(\tilde{\boldsymbol{\theta}})}{a(\psi)} + c(\mathbf{y}, \psi).$$

The latter is the highest possible value for the log-likelihood function achievable with the given data. The deviance function is then given by

$$\begin{aligned} D(\boldsymbol{\theta}, \mathbf{y}) &= 2 \sum_{i=1}^n [\ell(\tilde{\boldsymbol{\theta}}, \psi | \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}, \psi | \mathbf{y})] \\ &= 2 \sum_{i=1}^n [y_i(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) - b(\tilde{\boldsymbol{\theta}}) + b(\hat{\boldsymbol{\theta}})] a(\psi)^{-1}. \end{aligned}$$

This statistic is asymptotically chi-square with $n - k$ degrees of freedom (although high levels of discrete granularity in the outcome variable can make this a poor distributional assumption for data sets that are not so large). Fortunately, these deviance functions are commonly tabulated for many exponential family forms and therefore do not require analytical calculation. In the binomial case, the deviance function is

$$\begin{aligned} D(m, p) &= 2 \sum \left[y_i \log \left(\frac{y_i}{\mu_i} \right) \right. \\ &\quad \left. + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \mu_i} \right) \right], \end{aligned}$$

where the saturated log likelihood achieves the highest possible value for fitting $p_i = y_i/n_i$.

We can also look at the individual deviance contributions in an analogous way to linear model residuals. The single-point deviance function is just the deviance function for the y_i th point:

$$d(\boldsymbol{\theta}, y_i) = -2[y_i(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) - (b(\tilde{\boldsymbol{\theta}}) - b(\hat{\boldsymbol{\theta}}))] a(\psi)^{-1}.$$

A deviance residual at the point is built on this by

$$\mathbf{R}_{\text{deviance}} = \frac{(y_i - \mu_i)}{|y_i - \mu_i|} \sqrt{|d(\boldsymbol{\theta}, y_i)|},$$

where $\frac{(y_i - \mu_i)}{|y_i - \mu_i|}$ is just a sign-preserving function.

Table 2 Happiness and Schooling by Sex, 1977 General Social Survey

		Years of Schooling			
		< 12	12	13–16	17 +
Male	Not happy	40	21	14	3
	Pretty happy	131	116	112	27
	Very happy	82	61	55	27
Female	Not happy	62	26	12	3
	Pretty happy	155	156	95	15
	Very happy	87	127	76	15

EXAMPLE: MULTINOMIAL RESPONSE MODELS

Consider a slight generalization of the running binomial example in which, instead of two possible events defining the outcome variable, there are now k vents. The outcome for an individual i is given as a $k - 1$ length vector of all zeros—except for a single one identifying the positive response, $\mathbf{y}_i = [y_{i1}, y_{i2}, \dots, y_{i(k-1)}]$ —or all zeros for an individual selecting the left-out reference category. It should be clear that this reduces to a binomial outcome for $k = 2$.

The objective is now to estimate the $k - 1$ length of categorical probabilities for a sample size of n , $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\pi} = [\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots, \boldsymbol{\pi}_{k-1}]$, from the data set consisting of the $n \times (k - 1)$ outcome matrix $\mathbf{y}$ and the $n \times p$ matrix of covariates $\mathbf{X}$, including a leading column of 1s. The PMF for this setup is now multinomial, where the estimates are provided with a logit (but sometimes a probit) link function, giving for each of $k - 1$ categories the probability that the i th individual picks category r :

$$P(\mathbf{y}_i = r | \mathbf{X}) = \frac{\exp(\mathbf{X}_i \boldsymbol{\beta}_r)}{1 + \sum_{s=1}^{k-1} \exp(\mathbf{X}_i \boldsymbol{\beta}_s)},$$

where $\boldsymbol{\beta}_r$ is the coefficient vector for the r th category.

The data used are from the 1977 General Social Survey and are a specific three-category multinomial example analyzed by a number of authors. These are summarized in Table 2. Because there are only three categories in this application, the multinomial PMF for the i th individual is given by

$$f(\mathbf{y}_i | n_i, \boldsymbol{\pi}_i) = \frac{n_i}{y_{i1}! y_{i2}! (n - y_{i1} - y_{i2})!} \pi_{i1}^{y_{i1}} \pi_{i2}^{y_{i2}} (1 - \pi_{i1} - \pi_{i2})^{n - y_{i1} - y_{i2}},$$

and the likelihood is formed by the product across the sample. Is this an exponential family form such that

we can treat this as a GLM problem? This PMF can be reexpressed as

$$f(\mathbf{y}_i | n_i, \boldsymbol{\pi}_i) = \exp \left[\underbrace{\left(\frac{\mathbf{y}_i}{n_i}, \frac{\mathbf{y}_i}{n_i} \right)}_{\mathbf{y}'_i} \underbrace{\left(\log \left(\frac{\boldsymbol{\pi}_1}{1 - \boldsymbol{\pi}_1 - \boldsymbol{\pi}_2} \right), \log \left(\frac{\boldsymbol{\pi}_2}{1 - \boldsymbol{\pi}_1 - \boldsymbol{\pi}_2} \right) \right)'}_{\boldsymbol{\theta}_1} \right. \\ \left. - \underbrace{(-\log(1 - \boldsymbol{\pi}_1 - \boldsymbol{\pi}_2))}_{b(\boldsymbol{\theta}_2)} n_i + \log \left(\underbrace{\frac{n_i}{y_{i1}! y_{i2}! (n - y_{i1} - y_{i2})!}}_{c(\mathbf{y})} \right) \right],$$

where n_i is a weighting for the i th case. This is clearly an exponential form, albeit now with a two-dimensional structure for $\mathbf{y}'_i$ and $\boldsymbol{\theta}_i$. The two-dimensional link function that results from this form is simply

$$\boldsymbol{\theta}_i = g(\boldsymbol{\pi}_{i1}, \boldsymbol{\pi}_{i2}) \\ = \left(\log \left(\frac{\boldsymbol{\pi}_{i1}}{1 - \boldsymbol{\pi}_{i1} - \boldsymbol{\pi}_{i2}} \right), \log \left(\frac{\boldsymbol{\pi}_{i2}}{1 - \boldsymbol{\pi}_{i1} - \boldsymbol{\pi}_{i2}} \right) \right).$$

We can therefore interpret the results in the following way for a single respondent:

$$\log \left[\frac{P(\text{event 1})}{P(\text{reference category})} \right] = \log \left[\frac{\boldsymbol{\pi}_{i1}}{1 - \boldsymbol{\pi}_{i1} - \boldsymbol{\pi}_{i2}} \right] \\ = \mathbf{X}_i \boldsymbol{\beta}_1,$$

$$\log \left[\frac{P(\text{event 2})}{P(\text{reference category})} \right] = \log \left[\frac{\boldsymbol{\pi}_{i2}}{1 - \boldsymbol{\pi}_{i1} - \boldsymbol{\pi}_{i2}} \right] \\ = \mathbf{X}_i \boldsymbol{\beta}_2.$$

The estimated results for this model are contained in Table 3. What we observe from these results is that there is no evidence of gender effect (counter to other studies), and there is strong evidence of increased happiness for the three categories relative to the reference category of < 12 years of school. Interesting, the *very happy* to *not happy* distinction increases with

Table 3 Three-Category Multinomial Model Results

	Intercept	Years of Schooling			
		Female	12	13–16	17 +
Pretty	1.129	−0.181	0.736	1.036	0.882
happy	(0.148)	(0.168)	(0.196)	(0.238)	(0.453)
Very	0.474	0.055	0.878	1.114	1.451
happy	(0.161)	(0.177)	(0.206)	(0.249)	(0.455)

education, but the *pretty happy* to *not happy* distinction does not. The deviance residual is 8.68, indicating a good fit for 6 degrees of freedom (not in the tail of a chi-square distribution), and therefore an improvement over the saturated model.

—Jeff Gill

REFERENCES

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.
Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London, Series A*, 144, 285–307.
Gill, J. (2000). *Generalized linear models: A unified approach*. Thousand Oaks, CA: Sage.
McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York: Chapman & Hall.
Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135, 370–385.

GEOMETRIC DISTRIBUTION

A geometric distribution is used to model the number of failures ($n - 1$) of a binomial distribution before the first success (r). It is a special case of the NEGATIVE BINOMIAL DISTRIBUTION, or PASCAL DISTRIBUTION, when $r = 1$. The geometric distribution is a series of binomial trials. Due to the independence of binomial trials, the desired trial results are multiplied together to determine the overall PROBABILITY of the occurrence of the first success. For example, the geometric distribution would be used to model the number of tosses of a coin necessary to obtain the first “tails.”

A discrete RANDOM VARIABLE X is said to have a geometric distribution if it has a probability mass function in the form of

$$P(X = x) = p(1 - p)^{(x-1)}, \quad (1)$$

where

$$x = 1, 2, 3, \dots \quad (2)$$

and p = probability of success,

$$0 < p < 1. \quad (3)$$

For X to have a geometric distribution, each trial must have only two possible outcomes, success and failure. The outcomes of each trial must be statistically independent, and all trials must have the same probability of success over an infinite number of trials.

If X is a random variable from the geometric distribution, it has expectation

$$E(x) = \frac{1}{p} \quad (4)$$

and variance

$$\text{Var}(x) = \frac{(1 - p)}{p^2}. \quad (5)$$

The geometric distribution has a special property. It is “memoryless.” That means that the probability of observing a certain number of failures in a row does not depend on how many trials have gone before. So, the chance that the next five trials will all be failures after already seeing 100 failures is the same as if you had only seen 10 failures. The distribution “forgets” how many failures have already occurred.

The geometric distribution is analogous to the exponential distribution. The geometric distribution is “discrete” because you can only have an integer number of coin flips (i.e., you cannot flip a coin 1.34 times). The exponential distribution is continuous. Instead of the number of trials to success, it waits for the amount of time for the first success.

—Ryan Bakker

See also DISTRIBUTION

REFERENCES

Johnson, N. L., & Kotz, S. (1969). *Discrete distributions*. Boston: Houghton Mifflin.
King, G. (1998). *Unifying political methodology: The likelihood theory of statistical inference*. Ann Arbor: University of Michigan Press.

GIBBS SAMPLING

Gibbs sampling is an iterative Monte Carlo procedure that has proved to be a powerful statistical tool for approximating complex probability distributions that could not otherwise be constructed. The technique is a special case of a general set of procedures referred to as MARKOV CHAIN MONTE CARLO (MCMC) METHODS. One of the most important applications of Gibbs sampling has been to Bayesian inference problems that are complicated by the occurrence of missing data. Other useful applications in the social sciences have been the analysis of hierarchical linear models and the analysis of item response models. For the reader interested in some of the theory underlying MCMC methods, good overviews can be found in Gelman, Carlin, Stern, and Rubin (1995) and Schafer (1997).

A very simple example is useful in explaining the Gibbs sampler. Suppose one had knowledge of the statistical form of the joint probability distribution of two variables, say, x_1 and x_2 . Further suppose that one was primarily interested in the marginal distribution of each variable separately and/or the probability distribution of some function of the two variables (e.g., $x_1 - x_2$). In cases in which the mathematical form of the joint distribution is not complex (e.g., the two variables have a bivariate normal distribution), one can analytically construct the desired probability distributions. However, suppose that the joint distribution is quite nonstandard and analytic procedures are not possible. How does one proceed in this case? If one could draw a Monte Carlo—simulated sample of size N from the joint distribution, empirical relative frequency distributions could be constructed to estimate the distributions of interest. However, suppose it is very difficult to even sample from the joint distribution due to its complexity. Gibbs sampling to the rescue! In many problems, even when one cannot sample from a joint distribution, one can readily construct Monte Carlo methods for sampling from the full conditional distributions, for example, $P(x_1|x_2)$ and $P(x_2|x_1)$. The following iterative scheme is used in Gibbs sampling. Given a starting value for x_1 , sample a value for x_2 from $P(x_2|x_1)$. Given the generated x_2 , sample a new value for x_1 from $P(x_1|x_2)$. If we continue this iterative scheme for a “sufficiently” long burn-in period, all generated (x_1, x_2) pairs beyond this point can be treated as

samples from the desired bivariate distribution. By constructing relative frequency distributions separately for x_1 , x_2 , and $x_1 - x_2$, one can approximate the desired distributions.

In Bayesian analyses of incomplete data sets, the problem is to construct the posterior distribution of the unknown parameters (UP) using only the observed (O) data. When missing data (M) have occurred, the form of this posterior distribution can be very complex. However, if we let $x_1 = \text{UP}$, $x_2 = \text{M}$, and condition on O, one can readily approximate the desired posterior distribution and various marginal distributions using the Gibbs sampling scheme.

Unresolved issues in using Gibbs sampling are how to determine how long the burn-in period must be to ensure convergence and how to construct the sample given the proper burn-in period.

The BUGS software package (available for free from www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml) provides a powerful and relatively user-friendly program (and references) for performing a wide range of statistical analyses using Gibbs sampling.

—Alan L. Gross

REFERENCES

- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

GINI COEFFICIENT

The Gini coefficient is a measure of inequality of wealth or income within a population between 0 and 1. The coefficient is a measure of the graph that results from the Lorenz distribution of income or wealth across the full population (see INEQUALITY MEASUREMENT). The graph plots “percentage of ownership” against “percentile of population.” Perfect equality of distribution would be a straight line at 45 degrees. Common distributions have the shape represented in Figure 1. The Gini coefficient is the ratio of the area bounded by the curve and the 45-degree line divided by the area of the surface bounded by the 45-degree line and the x-axis.

—Daniel Little

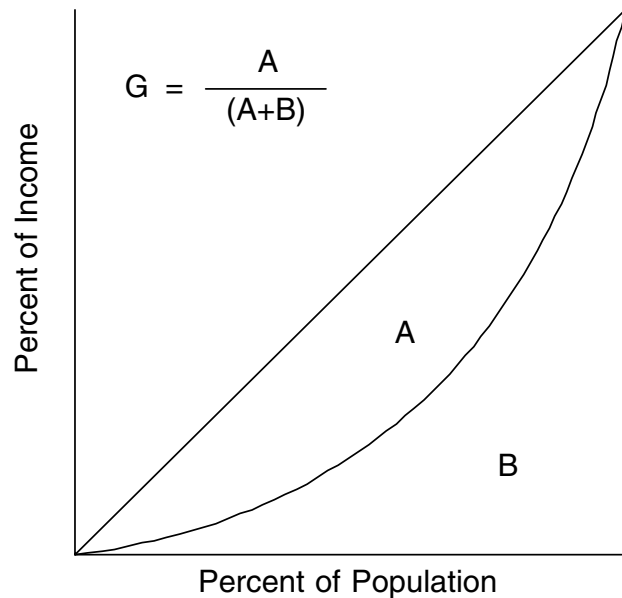


Figure 1 Lorenz Distribution of Income

GLIM

GLIM is a statistical software package for the estimation of GENERALIZED LINEAR MODELS. It is flexible in handling user-defined macros. For further information, see the Web site of the software at www.nag.co.uk/stats/GDGE_soft.asp.

—Tim Futing Liao

GOING NATIVE

At the more qualitative end of the research methods spectrum, the phrase “going native” encapsulates the development of a sense of “overrapport” between the researcher and those “under study,” to the extent that the researcher essentially “becomes” one of those under study. Seemingly imbued with negative connotations concerning the overembracing of aspects of “native” culture and ways of life by imperial and colonial ruling elites, its origins have often been attributed to Malinowski and his call for anthropologists to participate in the cultures they were observing to enhance their understanding of those cultures and peoples (Kanuha, 2000). Such calls for participation stood in stark contrast to more dominant expectations for any

(social) researcher to be the epitome of OBJECTIVITY. To be “committed,” to be *inside* the group, to work *with* the group would result in the wholesale adoption of an uncritical, unquestioning position of approval in relation to that group and their actions; the “standard” of the research would be questioned, its VALIDITY threatened (Stanley & Wise, 1993). The inclusion of the “researcher as person” within the research would be associated with an apparent inability for that researcher to “suitably” distance himself or herself from the events in which he or she was participating, ultimately undermining the “authority” of the voice of the “researcher as academic.” However, as Kanuha (2000) notes, such a perspective has been increasingly challenged as

the exclusive domain of academia once relegated only to white, male, heterosexual researchers has been integrated by scholars who are people of color, people from the barrio, and those whose ancestors were perhaps the native objects of their mentors and now colleagues. (p. 440)

Writing from “marginal” perspectives, the role of the researcher—the relationship between “us” and “them” and between the researcher and “researched-community”—has increasingly been deconstructed and problematized in recent years, leading ultimately to increased awareness of the need for critical, reflexive thought regarding the implications of any researcher’s positionality and situatedness. A key notion here is that “the personal is the political” (Stanley & Wise, 1993) and vice versa (Fuller, 1999), epitomizing how, for many researchers, personal, political, academic, and often activist identities are one and the same—many researchers are always or already natives in the research they are conducting. As such, they occupy a space in which the situatedness of their knowledges and their positionalities are constantly renegotiated and critically engaged with, a space that necessarily involves the removal of artificial boundaries between categories such as researcher, activist, teacher, and person and proposes instead movement between these various identities to facilitate engagement between and within them (Routledge, 1996). Here there is a continual questioning of researchers’ social location (in terms of class, gender, ethnicity, etc.), in addition to the physical location of their research, their disciplinary location, and their political position and personality (Robinson, 1994). In addition, such questioning is included and made transparent within any

commentary on the work conducted, highlighted as an integral (and unavoidable) part of the research process, thereby successfully avoiding any false and misleading presentation of the researcher (and research itself) as inert, detached, and neutral.

—Duncan Fuller

REFERENCES

- Fuller, D. (1999). Part of the action, or “going native”? Learning to cope with the politics of integration. *Area*, 31(3), 221–227.
- Kanuha, V. K. (2000). “Being” native versus “going native”: Conducting social work research as an insider. *Social Work*, 45(5), 439–447.
- Robinson, J. (1994). White women researching/representing “others”: From anti-apartheid to postcolonialism. In A. Blunt & G. Rose (Eds.), *Writing, women and space* (pp. 97–226). New York: Guilford.
- Routledge, P. (1996). The third space as critical engagement. *Antipode*, 28(4), 399–419.
- Stanley, L., & Wise, S. (1993). *Breaking out again: Feminist ontology and epistemology*. London: Routledge.

GOODNESS-OF-FIT MEASURES

Goodness-of-fit measures are STATISTICS calculated from a SAMPLE of data, and they measure the extent to which the sample data are consistent with the MODEL being considered. Of the goodness-of-fit measures, *R*-SQUARED, also denoted R^2 , is perhaps the most well known, and it serves here as an initial example. In a LINEAR REGRESSION analysis, the VARIABLE Y is approximated by a model, set equal to $\hat{Y}$, which is a linear combination of a set of k EXPLANATORY VARIABLES, $\hat{Y} = +\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$. The COEFFICIENTS in the model, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$, are estimated from the data so that the sum of the squared differences between Y and $\hat{Y}$ is as small as possible, also called the SUM OF SQUARED ERRORS (SSE), $SSE = \Sigma(Y - \hat{Y})^2$. R^2 gauges the goodness of fit between the observed variable Y and the regression model $\hat{Y}$ and is defined as $R^2 = 1 - \frac{SSE}{SST}$, where $SST = \Sigma(Y - \bar{Y})^2$, which is the sum of the squared deviations around the sample mean $\bar{Y}$. It ranges in value from 0, indicating extreme lack of fit, to 1, indicating perfect fit. It has universal appeal because it can be interpreted as the square of the

CORRELATION between Y and $\hat{Y}$ or as the proportion of the total variance in Y that is accounted for by the model.

ADDITIONAL EXAMPLES

Goodness-of-fit measures are not limited to regression models but may apply to any hypothetical statement about the study POPULATION. The following are three examples of models and corresponding goodness-of-fit measures.

The HYPOTHESIS that a variable X has a NORMAL DISTRIBUTION in the population is an example of a model. With a sample of observations on X , the goodness-of-fit statistic D can be calculated as a measure of how well the sample data conform to a normal distribution. In addition, the D statistic can be used in the KOLMOGOROV-SMIRNOV TEST to formally test the hypothesis that X is normally distributed in the population.

Another example of a model is the hypothesis that the mean, denoted μ , of the distribution of X is equal to 100, or $\mu_X = 100$. The t -STATISTIC can be calculated from a sample of data to measure the goodness of fit between the sample mean $\bar{X}$ and the hypothesized population mean μ_X . The t -statistic can then be used in a t -TEST to decide if the hypothesis $\mu_X = 100$ should be rejected based on the sample data.

One more example is the INDEPENDENCE model, which states that two categorical variables are unrelated in the population. The numbers expected in each category of a CONTINGENCY TABLE of the sample data, if the independence model is true, are called the “EXPECTED FREQUENCIES,” and the numbers actually observed in the sample are called the “OBSERVED FREQUENCIES.” The goodness of fit between the hypothesized independence model and the sample data is measured with the χ^2 (chi-square) statistic, which can then be used with the χ^2 test (CHI-SQUARE TEST) to decide if the independence model should be rejected.

HYPOTHESIS TESTS

The examples above illustrate two distinct classes of goodness-of-fit measures that serve different purposes. Each summarizes the consistency between sample data and a model, but R^2 is a descriptive statistic with a clear intuitive interpretation, whereas D , t , and χ^2

are INFERENTIAL STATISTICS that are less meaningful intrinsically. The inferential measures are valuable for conducting HYPOTHESIS TESTS about goodness of fit, principally because each has a known SAMPLING DISTRIBUTION that often has the same name as the statistic (e.g., t and χ^2). Even if a model accurately represents the population under study, a goodness-of-fit statistic calculated from a sample of data will rarely indicate a perfect fit with the model, and the extent of departure from perfect fit will vary from one sample to another. This variation is captured in the sampling distribution of a goodness-of-fit measure, which is known for D , t , and χ^2 . These statistics become interpretable as measures of fit when compared to their sampling distribution, so that if the sample statistic lies sufficiently far out in the tail of its sampling distribution, it suggests that the sample data are improbable if the model is true (i.e., a poor fit between data and model). Depending on the rejection criteria for the test and the SIGNIFICANCE LEVEL chosen, the sample statistic may warrant rejection of the proposed model.

REGRESSION MODELS

A goodness-of-fit statistic is generally applicable to a specific type of model. However, within a class of models, such as linear regression models, it is common to evaluate goodness of fit with several different measures. Following are examples of goodness-of-fit measures used in regression analysis. These fall into two categories: (a) scalar measures that summarize the overall fit of a model and (b) model comparison measures that compare the goodness of fit between two alternative models. Some of the measures described are applicable only to linear regression models; others apply to the broader class of regression models estimated with MAXIMUM LIKELIHOOD ESTIMATION (MLE). The latter includes linear and LOGISTIC REGRESSION models as well as other models with categorical or limited dependent variables.

Scalar Measures

Scalar measures summarize the overall fit of the model, with a single number. R^2 is the best example of this. However, R^2 does have limitations, and there are a number of alternative measures designed to overcome these. One shortcoming is that R^2 applies to linear models only. There is no exact counterpart in nonlinear models or models with categorical outcomes,

although there are several “PSEUDO- R^2 ” measures that are designed to simulate the properties of R^2 in these situations. A popular pseudo- R^2 measure is the maximum likelihood R^2 , denoted R_{ML}^2 . R_{ML}^2 would reproduce R^2 if applied to a linear regression model, but it can be applied more generally to all models that are estimated with MLE. For the model being evaluated, M_k , with k explanatory variables, there is an associated statistic L_{M_k} , which is the likelihood of observing the sample data given the PARAMETERS of model M_k , and another statistic L_{M_0} , which is the likelihood of observing the sample data given the null model with no explanatory variables, M_0 . R_{ML}^2 is calculated using the ratio of these two likelihoods, $R_{ML}^2 = 1 - [L_{M_0}/L_{M_k}]^2$. For categorical outcomes, a more intuitive alternative is the count R^2 , R_{count}^2 , which is the proportion of the sample observations that was correctly classified by the model. See Long (1997, pp. 102–108) and Menard (1995, pp. 22–37) for more discussion of alternative measures.

A second limitation of R^2 is that it cannot become smaller as more variables are added to a model. This gives the false impression that a model with more variables is a better fitting model. ADJUSTED R^2 , denoted R_{adj}^2 , is a preferred alternative that does not necessarily increase as more variables are added. It is calculated in such a way that it devalues R^2 when too many variables are included relative to the number of observations, $R_{adj}^2 = 1 - \frac{n-1}{n-p}(1 - R^2)$, with p parameters and n observations.

Akaike’s Information Criterion (AIC) is another popular scalar measure of fit. Like R_{adj}^2 , it takes into account the number of observations in the sample and the number of variables in the model and places a premium on models with a small number of parameters per observation. For a linear model with n observations and p parameters, AIC is defined as $AIC = n \ln \left[\frac{SSE}{n} \right] + 2p$. A more general formulation applies to models estimated with MLE. If L is the likelihood of the sample data given the proposed model, then $AIC = \frac{2 \ln L + 2p}{n}$. AIC is interpreted as the “per observation contribution” to the model, so smaller values are preferred, and it is useful for comparing different models from the same sample as well as the same model across different samples.

Model Comparison Measures

In addition to providing a summary measure of fit, scalar measures, as described above, can be valuable

for evaluating the overall fit of one model relative to another. Increments in R^2 , R^2_{adj} , and some pseudo- R^2 measures are included in a general class of model comparison measures that have a PROPORTIONATE REDUCTION OF ERROR (PRE) interpretation. When the fit of a multivariate linear model with k explanatory variables, M_k , is compared to the fit of the null model with no explanatory variables, M_0 , the inferential F -statistic, also called the F -RATIO, is directly related to R^2 and is reported automatically in most statistical software packages. The F -statistic has advantages over R^2 in that it serves as the basis for formally testing, with the F -test, the null hypothesis that the fit of a MULTIVARIATE model M_k is not a significant improvement over the null model M_0 . This is equivalent to testing the hypothesis that all of the slope coefficients in the model are equal to zero, that is, $\beta_1 = \beta_2 = \dots = \beta_k = 0$.

Analogous to the F -statistic for linear models is the model χ^2 statistic that applies to the larger class of models estimated with MLE. As the F -statistic is related to R^2 , the model χ^2 is related to R^2_{ML} , and as the F -statistic is the basis of the F -test, the model χ^2 is the basis of the LIKELIHOOD RATIO test of the null hypothesis that the fit of a multivariate MLE model, M_k , is not a significant improvement over the null model M_0 or, equivalently, $\beta_1 = \beta_2 = \dots = \beta_k = 0$. The model χ^2 is a function of the ratio of the two likelihoods described above, L_{M_k} and L_{M_0} , and is calculated as $\chi^2 = -2 \ln \left[\frac{L_{M_0}}{L_{M_k}} \right]$.

The F -test and the likelihood ratio test are more broadly applicable than the examples suggest. They can be used to test for improved goodness of fit between any two models, one having a subset of the variables contained in the other. They can test for improvement provided by sets of variables or by a single variable. For MLE models, alternatives to the likelihood ratio test are the Wald test and the Lagrange multiplier test. When testing for a significant contribution to goodness of fit from a single variable, t -tests can be used in linear models and MLE models.

Despite the reliance on hypothesis tests to compare the goodness of fit between two models, these tests need to be used cautiously. Because the probability of rejecting the null hypothesis increases with sample size, overconfidence in these methods when samples are large can lead to conclusions that variables make a significant contribution to fit when the effects may actually be small and unimportant. The Bayesian Information Criterion (BIC) is an alternative that is

Table 1 Strength of Evidence for Improved Goodness of Fit

Absolute Value of Difference, $ \text{BIC}_{M_1} - \text{BIC}_{M_2} $	Strength of Evidence
0 – 2	Weak
2 – 6	Positive
6 – 10	Strong
> 10	Very strong

not affected by sample size. BIC can be applied to a variety of statistics that gauge model improvement, and it offers a method for assessing the degree of improvement provided by one model relative to another. For example, when applied to the model χ^2 , which implicitly compares M_k to M_0 , $\text{BIC}_{M_k} - \text{BIC}_{M_0} = -\chi^2 + k \ln n$. In general, when comparing any two MLE models, M_1 and M_2 , with one not necessarily a subset of the other, $\text{BIC}_{M_1} - \text{BIC}_{M_2} \approx 2 \ln \left[\frac{L_{M_2}}{L_{M_1}} \right]$. Because the model with the more negative BIC is preferred, if $\text{BIC}_{M_1} - \text{BIC}_{M_2} < 0$, then the first model is preferred. If $\text{BIC}_{M_1} - \text{BIC}_{M_2} > 0$, then the second model is preferred. Raftery (1996) suggests guidelines, presented in Table 1, for evaluating model improvement based on the absolute value of the difference between two BICs. See Raftery (1996, pp. 133–139) and Long (1997, pp. 110–112) for more discussion of the BIC.

—Jani S. Little

REFERENCES

- Greene, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.
- Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.
- Raftery, A. E. (1996). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (Vol. 25, pp. 111–163). Oxford, UK: Basil Blackwell.

GRADE OF MEMBERSHIP MODELS

Grade of Membership (GoM) models produce a classification of a set of objects in the form of fuzzy partitions where persons may have partial membership in

more than one group. The advantage of GoM compared to discrete partition classification methods (e.g., CLUSTER ANALYSIS; LATENT CLASS ANALYSIS [LCA]) is that it generally produces fewer classes whose parametric relations to measured variables are easier to interpret. Indeed, LCA may be viewed as a special case of GoM in which classes are “crisp,” and persons belong only in one group (i.e., there is no within-class heterogeneity). In factor analysis, the model is constructed to use only second-order (i.e., covariance or correlation matrix) moments so it cannot describe nonnormal multivariate distributions that might arise if there are multiple distinct groups (subpopulations) comprising the total set of cases analyzed.

DESCRIPTION

GoM deals with a family of random variables, $\{Y_{ij}\}_{ij}$, representing measurements. Here i ranges over a set of objects $\mathcal{I}$ (in demography, objects are usually individuals), j ranges over measurements $\mathcal{J}$, and Y_{ij} takes values in a finite set $\mathcal{L}_j$, which, without loss of generality, may be considered as a set of natural numbers $\{1, \dots, L_j\}$. Every individual i is associated with a vector of random variables $(Y_{ij})_j$, with this vector containing all available information on the individual. The idea is to construct a convex space or “basis,” $\{(Y_{kj}^0)_j\}_k$, $k = 1, \dots, K$, in the space of random vectors $(Y_{ij})_j$ and, for every individual, coefficients of position on this “basis,” $(g_{ik})_k$, such that the equations $\Pr(Y_{ij} = l) = \sum_{k=1}^K g_{ik} \Pr(Y_{kj}^0 = l)$, together with constraints $g_{ik} \geq 0$, $\sum_k g_{ik} = 1$, hold. One might treat these as K fuzzy sets, or “pure profiles,” with a representative member ($g_{ik} = 1$ for some k) of the k th profile described by $(Y_{kj}^0)_j$. The interpretation of profiles is derived from a posteriori analysis of the coefficients $\lambda_{kjl} = \Pr(Y_{kj}^0 = l)$ produced by GoM.

The GoM ALGORITHM searches for g_{ik} and λ_{kjl} that maximize the likelihood $\prod_{i \in \mathcal{I}} \prod_{j \in \mathcal{J}} \prod_{l \in \mathcal{L}_j} (\sum_{k=1}^K g_{ik} \lambda_{kjl})^{y_{ijl}}$, with constraints $g_{ik} \geq 0$, $\sum_k g_{ik} = 1$, $\lambda_{kjl} \geq 0$, and $\sum_l \lambda_{kjl} = 1$, where y_{ijl} is 1 if the j th measurement on an i th individual gives outcome l and 0 otherwise. If no response is given, then all $y_{ijl} = 0$, and the variable is missing and drops from the likelihood. The λ_{kjl} determine the distribution of Y_{kj}^0 —that is, $\Pr(Y_{kj}^0 = l)$ —and the g_{ik} are estimates of the grades of membership. Likelihood estimates of g_{ik} are consistent (i.e., as both number of individuals and number of measurements tend to infinity, g_{ik} estimates converge

to true values). As in standard statistical models, the λ_{kjl} , averaged over i , are also statistically consistent.

EXAMPLE

In an analysis of the National Long-Term Care Survey (Manton & XiLiang, 2003), 27 different measures of function coded on binary (2) or quaternary (4) variables were selected. A GoM analysis was performed identifying six profiles ($K = 6$), whose position is defined in terms of the 27 observed variables by six vectors, λ_{kjl} . In this analysis, the first profile has no disability, and the sixth has a high likelihood of almost all functional disability measures. The other four types divide into two pairs of vectors, with one pair describing a graduation of impairment on instrumental activities of daily living and a second pair describing a graduation of activities of daily living impairments. Because there is significant variation in the g_{ik} , no six classes of LCA could describe as much variability as GoM. Because the distribution of the g_{ik} s was not normal, no factor analysis (FA) could have fully described the data (including higher order moments) in as few as six dimensions.

HISTORY AND BIBLIOGRAPHY

The GoM model was first described in 1974 (Woodbury & Clive, 1974). Books (Manton & Stallard, 1988; Manton, Woodbury, & Tolley, 1994) provide a detailed description of the method and give examples of applications. Papers (Tolley & Manton, 1992) also contain proofs of the important statistical properties (e.g., sufficiency and consistency) of the g_{ik} and λ_{kjl} parameter estimates.

—Kenneth Manton and Mikhail Kovtun

REFERENCES

- Manton, K. G., & Stallard, E. (1988). *Chronic disease modelling: Vol. 2. Mathematics in medicine*. New York: Oxford University Press.
- Manton, K. G., & XiLiang, G. (2003). *Variation in disability decline and Medicare expenditures*. Working Monograph, Center for Demographic Studies.
- Manton, K. G., Woodbury, M. A., & Tolley, H. D. (1994). *Statistical applications using fuzzy sets*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley.
- Tolley, H. D., Kovtun, M., Manton, K. G., & Akushevich, I. (2003). Statistical properties of grade of membership

models for categorical data. Manuscript to be submitted to *Proceedings of National Academy of Sciences*, 2003.

Tolley, H. D., & Manton, K. G. (1992). Large sample properties of estimates of discrete grade of membership model. *Annals of the Institute of Statistical Mathematics*, 41, 85–95.

Woodbury, M. A., & Clive, J. (1974). Clinical pure types as a fuzzy partition. *Journal of Cybernetics*, 4(3), 111–121.

GRANGER CAUSALITY

Econometrician Clive Granger (1969) has developed a widely used definition of CAUSALITY. A key aspect of the concept of *Granger causality* is that there is no instantaneous causation (i.e., all causal processes involve time lags). X_t “Granger causes” Y_t if information regarding the history of X_t can improve the ability to predict Y_t beyond a prediction made using the history of Y_t alone. Granger causality is not unidirectional; that is, Y_t Granger causes X_t if information about the history of Y_t can improve the ability to predict X_t beyond a prediction made using only the history of X_t .

Granger causality tests may be performed using a bivariate vector autoregression (VAR) model. A VAR for X and Y specifies that each variable is a function of k lags of itself plus k lags of the other variable; that is,

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_k Y_{t-k} + \lambda_1 X_{t-1} + \cdots + \lambda_k X_{t-k} + \varepsilon_t, \quad (1)$$

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + \cdots + \alpha_k X_{t-k} + \gamma_1 Y_{t-1} + \cdots + \gamma_k Y_{t-k} + \zeta_t. \quad (2)$$

An F -test or a Lagrange multiplier (LM) test can be used to determine if X Granger causes Y or Y Granger causes X . For example, the LM test to see if X Granger causes Y involves estimating (1) under the assumption that the λ s are jointly zero. Residuals from this regression then are regressed on all variables in (1). The R^2 from this second regression $\times T$ (number of observations) is the test statistic (distributed as χ^2 , $df = 2k + 1$). Rejection of the null hypothesis implies X Granger causes Y . An alternative VAR representation of the above test has been developed by Sims (1972). The impact of nonstationarity on the validity of these tests is controversial. Granger causality tests also may be implemented in the ARIMA modeling framework (see Freeman, 1983).

The concept of Granger causality has been incorporated into the definitions of exogeneity offered by

Hendry and associates (e.g., Charemza & Deadman, 1997, chap. 7). Recognize that one need not establish that X Granger causes Y to warrant inference concerning the effects of X in a single-equation model of Y . It is sufficient to establish that X is *weakly exogenous* to Y (i.e., parameters in the model for Y can be efficiently estimated without any information in a model for X). However, if one wishes to forecast Y , then it is necessary to investigate if X is *strongly exogenous* to Y . Strong exogeneity equals weak exogeneity plus Granger “noncausality” (i.e., Y_t does not affect X at time t or at any time $t + k$). If Y is not strongly exogenous to X , then feedback from Y to X must be taken into account should that feedback occur within the forecast horizon of interest.

—Harold D. Clarke

REFERENCES

- Charemza, W. W., & Deadman, D. F. (1997). *New directions in econometric practice* (2nd ed.). Cheltenham, UK: Edward Elgar.
- Freeman, J. F. (1983). Granger causality and the time series analysis of political relationships. *American Journal of Political Science*, 27, 325–355.
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 24–36.
- Sims, C. A. (1972). Money income and causality. *American Economic Review*, 62, 540–552.

GRAPHICAL MODELING

A graphical model is a parametric statistical model for a set of random variables. Under this model, associations between the variables can be displayed using a mathematical graph. Graphical modeling is the process of choosing which graph best represents the data.

The use of mathematical graphs in statistics dates back to the PATH DIAGRAMS of Sewall Wright in the 1920s and 1930s. However, it was not until the seminal paper of John N. Darroch, Steffen L. Lauritzen, and Terry P. Speed, published in 1980, that a way of constructing a graph that has a well-defined PROBABILISTIC interpretation was proposed. This graph has since been called the *conditional independence graph* or *independence graph*, for short.

The (conditional) independence structure between variables X_1 to X_p can be displayed using a

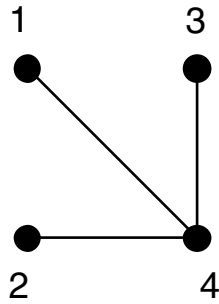


Figure 1 An Independence Graph

mathematical graph with p vertices (dots) corresponding to the p variables in which an edge (line) between vertices i and j is missing if, and only if, X_i and X_j are conditionally independent given the remaining $p - 2$ variables. For example, if X_1 to X_4 satisfy the following conditional independencies— X_1 independent of X_2 given X_3 and X_4 , X_1 independent of X_3 given X_2 and X_4 , and X_2 independent of X_3 given X_1 and X_4 —then their independence graph, displayed in Figure 1, has four vertices and edges missing between the following pairs of vertices: (1, 2), (1, 3), and (2, 3).

Although any set of jointly distributed random variables has an independence graph, the analyst requires, to construct the graph for some given data, a family of models in which conditional independence can easily be parameterized. For CONTINUOUS VARIABLES, multivariate normal distributions are used; for DISCRETE variables, LOG-LINEAR MODELS are used; and for a mixture of discrete and continuous variables, conditional Gaussian distributions are used. When there is a mixture, it is usual to indicate discrete variables by dots and continuous variables by circles in the graph.

When searching for a well-fitting graphical model, it is usual to first calculate the edge exclusion deviances: the LIKELIHOOD RATIO STATISTICS for the tests that each edge, in turn, is missing from the graph. The deviances are then compared with a preselected critical level to obtain a list of significant edges, which must be retained in the graph. The nonsignificant edges are then considered as edges, which could possibly be omitted from the graph, and model selection continues using procedures similar to those used in regression for selecting the best subset of the regressors. Once a well-fitting graph has been selected, information about the associations between the variables can be obtained from the graph using theoretical results, called the Markov properties.

Latent variable models can be represented by including vertices for the unobserved variables. So can Bayesian models by including vertices for the parameters. Directed graphs with arrows, rather than lines, between the vertices or chain graphs with a mixture of lines and arrows can be used to represent (potentially) causal relationships between variables or to distinguish between responses and explanatory variables.

—Peter W. F. Smith

REFERENCES

- Darroch, J. N., Lauritzen, S. L., & Speed, T. P. (1980). Markov fields and log-linear models. *Annals of Mathematical Statistics*, 43, 1470–1480.
- Edwards, D. (2000). *Introduction to graphical modelling* (2nd ed.). New York: Springer-Verlag.
- Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Chichester, UK: Wiley.

GROUNDING THEORY

Grounded theory refers to a set of systematic inductive methods for conducting QUALITATIVE RESEARCH aimed toward theory development. The term *grounded theory* denotes dual referents: (a) a *method* consisting of flexible methodological strategies and (b) the *products* of this type of inquiry. Increasingly, researchers use the term to mean the methods of inquiry for collecting and, in particular, analyzing data. The methodological strategies of grounded theory are aimed to construct middle-level theories directly from data analysis. The inductive theoretical thrust of these methods is central to their logic. The resulting analyses build their power on strong empirical foundations. These analyses provide focused, abstract, conceptual theories that explain the studied empirical phenomena.

Grounded theory has considerable significance because it (a) provides explicit, sequential guidelines for conducting qualitative research; (b) offers specific strategies for handling the analytic phases of inquiry; (c) streamlines and integrates data collection and analysis; (d) advances conceptual analysis of qualitative data; and (e) legitimizes qualitative research as scientific inquiry. Grounded theory methods have earned their place as a standard social research method and have influenced researchers from varied disciplines and professions.

Yet grounded theory continues to be a misunderstood method, although many researchers purport to use it. Qualitative researchers often claim to conduct grounded theory studies without fully understanding or adopting its distinctive guidelines. They may employ one or two of the strategies or mistake qualitative analysis for grounded theory. Conversely, other researchers employ grounded theory methods in reductionist, mechanistic ways. Neither approach embodies the flexible yet systematic mode of inquiry, directed but open-ended analysis, and imaginative theorizing from empirical data that grounded theory methods can foster. Subsequently, the potential of grounded theory methods for generating middle-range theory has not been fully realized.

DEVELOPMENT OF THE GROUNDED THEORY METHOD

The originators of grounded theory, Barney G. Glaser and Anselm L. Strauss (1967), sought to develop a systematic set of procedures to analyze qualitative data. They intended to construct theoretical analyses of social processes that offered abstract understandings of them. Their first statement of the method, *The Discovery of Grounded Theory* (1967), provided a powerful argument that legitimized qualitative research as a credible methodological approach, rather than simply as a precursor for developing quantitative instruments. At that time, reliance on quantification with its roots in a narrow POSITIVISM dominated and diminished the significance of qualitative research. Glaser and Strauss proposed that systematic qualitative analysis had its own logic and could generate theory. Their work contributed to revitalizing qualitative research and maintaining the ethnographic traditions of Chicago school sociology. They intended to move qualitative inquiry beyond descriptive studies into the realm of explanatory theoretical frameworks, thereby providing abstract, conceptual understanding of the studied phenomena. For them, this abstract understanding contrasted with armchair and logico-deductive theorizing because a grounded theory contained the following characteristics: a close fit with the data, usefulness, density, durability, modifiability, and explanatory power (Glaser, 1978, 1992; Glaser & Strauss, 1967).

Grounded theory combined Strauss's Chicago school intellectual roots in pragmatism and symbolic interactionism with Glaser's rigorous quantitative methodological training at Columbia University with

Paul Lazarsfeld. Strauss brought notions of agency, emergence, meaning, and the pragmatist study of action to grounded theory. Glaser employed his analytic skills to codify qualitative analysis. They shared a keen interest in studying social processes. They developed a means of generating substantive theory by invoking methodological strategies to explicate fundamental social or social psychological processes within a social setting or a particular experience such as having a chronic illness. The subsequent grounded theory aimed to understand the studied process, demonstrate the causes and conditions under which it emerged and varied, and explain its consequences. Glaser and Strauss's logic led them to formal theorizing as they studied generic processes across substantive areas and refined their emerging theories through seeking relevant data in varied settings.

Although the *Discovery of Grounded Theory* transformed methodological debates and inspired generations of qualitative researchers, Glaser's book, *Theoretical Sensitivity* (1978), provided the most definitive early statement of the method. Since then, Glaser and Strauss have taken grounded theory in somewhat separate directions (Charmaz, 2000). Glaser has remained consistent with his earlier exegesis of the method, which relied on direct and, often, narrow EMPIRICISM. Strauss has moved the method toward verification, and his coauthored works with Juliet M. Corbin (Strauss & Corbin, 1990, 1998) furthered this direction.

Strauss and Corbin's (1990, 1998) version of grounded theory also favors their new techniques rather than emphasizing the comparative methods that distinguished earlier grounded theory strategies. As before, Glaser studies basic social processes, uses comparative methods, and constructs abstract relationships between theoretical categories. Glaser's position assumes an external reality independent of the observer, a neutral observer, and the discovery of data. Both approaches have strong elements of objectivist inquiry founded in positivism. However, Glaser (1992) contends that Strauss and Corbin's procedures force data and analysis into preconceived categories and thus contradict fundamental tenets of grounded theory. Glaser also states that study participants will tell researchers what is most significant in their setting, but participants often take fundamental processes for granted. Showing how grounded theory can advance constructionist, interpretative inquiry minimizes its positivist cast.

Grounded theory has gained reknown because its systematic strategies aid in managing and analyzing qualitative data. Its REALIST cast, rigor, and positivistic assumptions have made it acceptable to quantitative researchers. Its flexibility and legitimacy appeal to qualitative researchers with varied theoretical and substantive interests.

APPLYING GROUNDED THEORY METHODS

Grounded theory methods provide guidelines for grappling with data that stimulate ways of thinking about them. These methods are fundamentally comparative. By comparing data with data, data with theoretical categories, and category with category, the researcher gains new ideas (Glaser, 1992). Discoveries in grounded theory emerge from the depth of researchers' empirical knowledge and their analytic interaction with the data; discoveries do not inhere in the data. Although researchers' disciplinary perspectives sensitize them to conceptual issues, grounded theory methods prompt researchers to start but not end with these perspectives.

These methods involve researchers in early data analysis from the beginning of data collection. Early analytic work then directs grounded theory researchers' subsequent data collection. Grounded theory strategies keep researchers actively engaged in analytic steps. These strategies also contain checks to ensure that the researcher's emerging theory is grounded in the data that it attempts to explain. Which data researchers collect and what they *do* with them shapes how and what they think about them. In turn, what they think about their data, including the questions they ask about them, guide their next methodological step. Moreover, grounded theory methods shape the content of the data as well as the form of analysis.

Grounded theory directs researchers to take a progressively analytic, methodical stance toward their work. They increasingly focus data collection to answer specific analytic questions and to fill gaps in the emerging analysis. As a result, they seek data that permit them to gain a full view of the studied phenomena. Grounded theory strategies foster gathering rich data about specific processes rather than the general structure of one social setting. These strategies also foster making the analysis progressively more focused and abstract. Subsequently, a grounded theory analysis may provide a telling explanation of the studied phenomena from which other researchers may

deduce hypotheses. Because grounded theory methods aim to generate theory, the emerging analyses consist of abstract concepts and their interrelationships. The following grounded theory strategies remain central to its logic as they cut across all variants of the method.

CODING DATA

The analytic process begins with two phases of coding: open, or initial, and selective, or focused. In open coding, researchers treat data analytically and discover what they see in them. They scrutinize the data through analyzing and coding each line of text (see Table 1, for example).

This line-by-line coding keeps researchers open to fresh ideas and thus reduces their inclinations to force the data into preconceived categories. Such active analytic involvement fosters developing theoretical sensitivity because researchers begin to ask analytic questions of the data. Grounded theorists ask a fundamental question: What is happening here? But rather than merely describing action, they seek to identify its phases, preconditions, properties, and purposes. Glaser and Strauss' early statements specify analyzing the basic social or social psychological process in the setting. Studying the fundamental process makes for a compelling analysis. However, multiple significant processes may be occurring; the researcher may study several of them simultaneously and demonstrate their interconnections. In a larger sense, grounded theory methods foster studying any topic as a process.

This initial coding forces researchers to think about small bits of data and to interpret them—immediately. As depicted in Table 1, grounded theory codes should be active, specific, and short.

Through coding, researchers take a new perspective on the data and view them from varied angles. Line-by-line coding loosens whatever taken-for-granted assumptions that researchers may have shared with study participants and prompts taking their meanings, actions, and words as problematic objects of study. This close scrutiny of data prepares researchers to compare pieces of data. Thus, they compare statements or actions from one individual, different incidents, experiences of varied participants, and similar events at different points in time. Note, for example, that the interview participant in Table 1 mentions that her friend and her husband provide different kinds of support. The researcher can follow the leads in her statements

Table 1 Example of Initial Coding

<i>Line-by-Line Coding</i>	<i>Interview Statement of a 68-Year-Old Woman With Chronic Illness</i>
Reciprocal supporting	We're [a friend who has multiple sclerosis] kind of like mutual supporters for each other. And
Having bad days	when she has her bad days or when we
Disallowing self-pity	particularly feel "poor me," you know, "Get off
Issuing reciprocal orders	your butt!" You know, we can be really pushy to
Comparing supportive others	each other and understand it. But with Fred [husband who has diabetes and heart disease]
Defining support role	here to support me no matter what I wanted to do.
Taking other into account	. . . And he's the reason why I try not to think of
Interpreting the unstated	myself a lot . . . because he worries. If he doesn't
Invoking measures of health/illness	see me in a normal pattern, I get these side
Averting despair	glances, you know, like "Is she all right?" And so
Transforming suffering	he'll snap me out of things and we're good for
Acknowledging other's utmost significance	each other because we both bolster, we need each other. And I can't fathom life without him. I know I can survive, you know, but I'd rather not.

to make a comparative analysis of support and the conditions under which it is helpful.

When doing line-by-line coding, researchers study their data, make connections between seemingly disparate data, and gain a handle on basic processes. Then they direct subsequent data collection to explore the most salient processes. Line-by-line coding works particularly well with transcriptions of interviews and meetings. Depending on the level and quality of field note descriptions, it may not work as well with ethnographic material. Hence, researchers may code larger units such as paragraphs or descriptions of anecdotes and events.

Initial coding guides focused or selective coding. After examining the initial codes, researchers look for both the most frequent codes and those that provide the most incisive conceptual handles on the data. They adopt these codes for focused coding. Focused codes provide analytic tools for handling large amounts of data; reevaluating earlier data for implicit meanings, statements, and actions; and, subsequently, generating categories in the emerging theory. Increasingly, researchers use computer-assisted programs for coding and managing data efficiently. Whether such programs foster reductionist, mechanistic analyses remains a question.

MEMO WRITING

Memo writing is the pivotal intermediate stage between coding and writing the first draft of the report.

Through writing memos, researchers develop their analyses in narrative form and move descriptive material into analytic statements. In the early stages of research, memo writing consists of examining initial codes, defining them, taking them apart, and looking for assumptions, meanings, and actions imbedded in them as well as relationships between codes. Early memos are useful to discuss hunches, note ambiguities, raise questions, clarify ideas, and compare data with data.

These memos also advance the analysis because they prompt researchers to explicate which comparisons between data are the most telling and how and why they are. Such comparisons help researchers to define fundamental aspects of participants' worlds and actions and to view them as processes. Subsequently, they can decide which processes are most significant to study and then explore and explicate them in memos. Writing memos becomes a way of studying data. Researchers can then pursue the most interesting leads in their data through further data collection and analysis.

Memo writing enables researchers to define and delineate their emerging theoretical categories. Then they write later memos to develop these categories and to ascertain how they fit together. They compare data with the relevant category, as well as category to category. Writing progressively more theoretical memos during their studies increases researchers' analytic competence in interpreting the data as well as their confidence in the interpretations.

By making memos increasingly more analytic with full coverage of each category, researchers gain a handle on how they fit into the studied process. Then they minimize what can pose a major obstacle: how to integrate the analysis into a coherent framework.

THEORETICAL SAMPLING

As researchers write memos explicating their codes and categories, they discern gaps that require further data collection. Subsequently, they engage in theoretical sampling, which means gathering further, more specific data to illuminate, extend, or refine theoretical categories and make them more precise. Thus, the researcher's emerging theory directs this sampling, not the proportional representation of population traits. With skillful use of this sampling, researchers can fill gaps in their emerging theoretical categories, answer unresolved questions, clarify conditions under which the category holds, and explicate consequences of the category.

Theoretical sampling may be conducted either within or beyond the same field setting. For example, researchers may return to the field and seek new data by posing deeper or different research questions, or they may seek new participants or settings with similar characteristics to gain a fresh view. However, as researchers' emerging theories address generic processes, they may move across substantive fields to study their concepts and develop a formal theory of the generic process. Through theoretical sampling, researchers make their categories more incisive, their memos more useful, and their study more firmly grounded in the data.

CONSTRUCTING AND LINKING THEORETICAL CATEGORIES

Coding, memo writing, and theoretical sampling progressively lead researchers to construct theoretical categories that explain the studied process or phenomena. Some categories arise directly from open coding, such as research participants' telling statements. Researchers successively infer other categories as their analyses become more abstract and subsume a greater range of codes.

Grounded theorists invoke the saturation of core categories as the criterion for ending data collection. But what does that mean? And whose definition of saturation should be adopted? If the categories are

concrete and limited, then saturation occurs quickly. Some studies that purport to be grounded theories are neither well grounded nor theoretical; their categories lack explanatory power, and the relationships between them are either unimportant or already established.

For a grounded theory to have explanatory power, its theoretical categories should be abstract, explicit, and integrated with other categories. Thus, grounded theorists complete the following tasks: locating the context(s) in which the category is relevant; defining each category; delineating its fundamental properties; specifying the conditions under which the category exists or changes; designating where, when, and how the category is related to other categories; and identifying the consequences of these relationships. These analytic endeavors produce a fresh theoretical understanding of the studied process.

—Kathy Charmaz

REFERENCES

- Charmaz, K. (2000). Grounded theory: Constructivist and objectivist methods. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 509–535). Thousand Oaks, CA: Sage.
- Glaser, B. (1978). *Theoretical sensitivity*. Mill Valley, CA: Sociology Press.
- Glaser, B. (1992). *Emergence vs. forcing: Basics of grounded theory analysis*. Mill Valley, CA: Sociology Press.
- Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Chicago: Aldine.
- Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.
- Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Grounded theory procedures and techniques* (2nd ed.). Thousand Oaks, CA: Sage.

GROUP INTERVIEW

Group interviewing is the systematic questioning of several individuals simultaneously in formal or informal settings (Fontana & Frey, 1994). This technique has had a limited use in social science, where the primary emphasis has been on interviewing individuals. The most popular use of the group interview has been in FOCUS GROUPS, but there are additional formats for this technique, particularly in qualitative fieldwork. Depending on the research purpose, the

interviewer/researcher can direct group interaction and response in a structured or unstructured manner in formal or naturally occurring settings. Groups can be used in exploratory research to test a research idea, survey question, or a measurement technique. Brainstorming with groups has been used to obtain the best explanation for a certain behavior. Group discussion will also enable researchers to identify key respondents (Frey & Fontana, 1991). This technique can be used in combination with other data-gathering strategies to obtain multiple views of the same research problem. The phenomenological assessment of emergent meanings that go beyond individual interpretation is another dimension of group interviews.

Group interviews can have limited structure (e.g., direction from interviewer) as in the case with brainstorming or exploratory research, or there can be considerable structure as in the case with nominal, Delphi, or most marketing focus groups. Respondents can be brought to a formal setting, such as a laboratory in a research center, or the group can be stimulated in an informal field setting, such as a street corner, recreation center, or neighborhood gathering place. Any group interview, regardless of setting, requires significant ability on the part of the interviewer to recognize an interview opportunity, to ask the questions that stimulate needed response, and to obtain a representative response or one that does not reflect the view of the dominant few and neglect perspectives of otherwise timid participants. The interviewer must be able to draw balance between obtaining responses related to the research purpose and the emergent group responses. Thus, the interviewer can exercise considerable control over the questioning process using a predetermined script of questions, or the interviewer can exercise little or no control over the direction and content of response by simply letting the conversation and response take their own path.

Problems exist when appropriate settings for group interviews cannot be located or when such a setting (e.g., public building, street corner) does not lend itself to an expressive, often conflict-laden, discussion. It is also possible that the researcher will not be accepted by the group as either a member or researcher. The group members may also tend to conformity rather than diversity in their views, thus not capturing the latent variation of responses and making it difficult to approach sensitive topics (Frey & Fontana, 1991). Finally, there is a high rate of irrelevant data; recording and compiling field notes is problematic,

particularly in natural, informal settings; and posturing by respondents is more likely in the group.

Advantages include efficiency, with the opportunity to get several responses at the same time and in the same settings. Group interviews provide another and rich dimension to understanding an event or behavior; respondents can be stimulated by the group to recall events they may have forgotten, and the researcher gains insights into social relationships associated with certain field settings.

—James H. Frey

REFERENCES

- Fontana, A., & Frey, J. H. (1994). Interviewing: The art of science. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 361–376). Thousand Oaks, CA: Sage.
- Fontana, A., & Frey, J. H. (2000). The interview: From structured questions to negotiated text. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 645–672). Thousand Oaks, CA: Sage.
- Frey, J. H., & Fontana, A. (1991). The group interview in social research. *The Social Science Journal*, 28, 175–187.

GROUPED DATA

Data come from VARIABLES measured on one or more units. In the social sciences, the unit is often an individual. Units used in other sciences could be elements such as a pig, a car, or whatever. It is also possible to have *larger* units, such as a county or a nation. Such units are often made up by aggregating data across individual units. We aggregate the values of a variable across the individual persons and get *grouped data* for the larger unit. Grouped data can consist of data on variables such as average income in a city or number of votes cast in a ward for a candidate.

However, data on aggregates do not always consist of grouped data. If we consider type of government in a country as a variable with values *Democracy*, *Dictatorship*, and *Other*, then we clearly have data on aggregates of individuals in the countries. But the value for any country is a characteristic of the people who make up the country, and it is not an aggregate of values of any variable across the individuals in that country.

Grouped data exist at various levels of aggregation. There can be grouped data for the individuals who

make up a census tract, a country, or a state. The importance of this lies in the fact that aggregation to different levels can produce different results from analyses of the data. Data on different levels may produce different magnitudes of CORRELATIONS depending on whether the data consist of observations aggregated, say, to the level of the county, where we may have data on all 3,000+ counties in the country, or whether the data are aggregated to the level of the state, where we may have data on all 50 states. Thus, analysis results that come from one level of aggregation apply *only* to the level on which the analysis is done. Mostly, they do not apply to units at lower or higher levels of aggregation.

In particular, results obtained from the analysis of aggregate data do in no way necessarily apply to the level of individuals as well. The so-called ECOLOGICAL FALLACY occurs when results obtained from grouped data are thought to apply on the level of the individual as well. In sociology, Robinson (1950) made this very clear in his path-breaking article on this topic. He showed mathematically how an ecological correlation coefficient obtained from grouped data could be very different, both in magnitude and in sign, from the correlation of the same two variables using data on individuals. *Simpson's paradox* is another name used for this phenomenon.

This situation becomes worse when there are group data available, but the individual data have not been observed. In voting, we know the number of votes cast for a candidate as well as other characteristics of precincts, but we do not know how the single individuals voted. Attempts have been made to construct methods to recover underlying individual-level data from group data, but in principle, such recovery will remain impossible without additional information.

In social science data analysis, another common form of grouped data is cross-classified data (i.e., observations cross-classified by the categories of the variables in an analysis). Such cross-classifications are known as CONTINGENCY TABLES and are often analyzed by LOG-LINEAR MODEL.

—Gudmund R. Iversen

See also DATA

REFERENCES

- Borgatta, E. F., & Jackson, D. J. (Eds.). (1980). *Aggregate data: Analysis and interpretation*. Beverly Hills, CA: Sage.
- Iversen, G. R. (1973). Recovering individual data in the presence of group and individual effects. *American Journal of Sociology*, 79, 420–434.
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351–357.

GROWTH CURVE MODEL

Growth curve models refer to a class of techniques that analyze trajectories of cases over time. As such, they apply to longitudinal or PANEL data, where the same cases are repeatedly observed. John Wishart (1938) provided an early application of the growth curve model in which he fit polynomial growth curves to analyze the weight gain of individual pigs. Zvi Griliches (1957) looked at the growth of hybrid corn across various regions in the United States and modeled predictors of these growth parameters. Since these early works, growth curve models have increased in their generality and have spread across numerous application areas.

The word *growth* in growth curve models reflects the origin of these procedures in the biological sciences, whereby the organisms studied typically grew over time, and a separate growth trajectory could be fit to each organism. With the spread of these techniques to the social and behavioral sciences, the term *growth* seems less appropriate, and there is some tendency to refer to these models as latent curve models or *latent trajectory models*. We will use the term latent curve models in recognition that the objects of study might grow, decline, or follow other patterns rather than linear growth.

Social and behavioral scientists approach latent curve models from two methodological perspectives. One treats latent curve models as a special case of multilevel (or hierarchical) models, and the other treats them as a special case of STRUCTURAL EQUATION MODELS (SEMs). Although these methodological approaches have differences, the models and estimators derived from multilevel models or SEMs are sometimes identical and often are similar. A basic distinction that holds across these methodological approaches is whether a latent curve model is unconditional or conditional. The next two sections describe these two types of models.

UNCONDITIONAL MODEL

The unconditional latent curve model refers to REPEATED MEASURES for a single outcome and the modeling of this variable's trajectory. We can conceptualize the model as consisting of Level 1 and Level 2 equations. The Level 1 equation is

$$y_{it} = \alpha_i + \beta_i \lambda_t + \varepsilon_{it},$$

where i indexes the case or observation, and t indexes the time or wave of the data. The y_{it} is the outcome or repeated-measure variable for the i th case at the t th time point, α_i is the intercept for the i th case, β_i is the slope for the i th case, and λ_t is a function of the time of observation. The typical assumption is that $\lambda_t = t - 1$. In this case, the model assumes that the outcome variable is adequately captured by a linear trend. Nonlinear trends are available, and we will briefly mention these in our last section on extensions of the model. Finally, ε_{it} is the random DISTURBANCE for the i th case at the t th time period, which has a mean of zero ($E(\varepsilon_{it}) = 0$) and is uncorrelated with α_i , β_i , and λ_t . In essence, the Level 1 equation assumes an individual trajectory for each case in the sample, where the intercept (α_i) and slope (β_i) determine the trajectory and can differ by case. This part of the model differs from many typical social science statistical models in that each individual is permitted to follow a different trajectory.

The Level 2 equations treat the intercepts and slopes as "dependent" variables. In the unconditional model, the Level 2 equations give the group mean intercept and mean slope so that

$$\alpha_i = \mu_\alpha + \zeta_{\alpha i},$$

$$\beta_i = \mu_\beta + \zeta_{\beta i},$$

where μ_α is the mean of the intercepts, μ_β is the mean of the slopes, $\zeta_{\alpha i}$ is the disturbance that is the deviation of the intercept from the mean for the i th case, and $\zeta_{\beta i}$ is a similarly defined disturbance for β_i . The disturbances have means of zero, and they can correlate with each other. Generally, researchers need at least three waves of data to identify all of the model parameters.

The preceding Level 1 and Level 2 equations have the same form in both the multilevel and the SEM approaches to these models. In the SEM tradition of latent curve models, a path diagram is an alternative way to present these models. Figure 1 is a path diagram of the Level 1 and Level 2 equations for an unconditional latent curve model with four waves of data. In

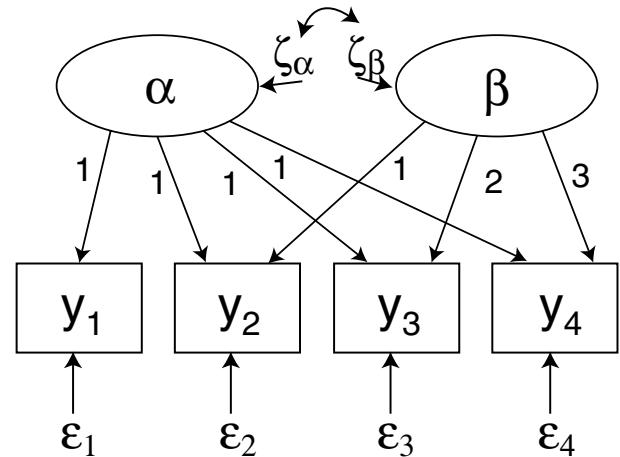


Figure 1 Unconditional Latent Trajectory Model

this path diagram, circles (or ellipses) represent latent variables, observed variables are in squares or rectangles, and disturbances are not enclosed. Single-headed arrows stand for the impact of the variable at the base of the arrow on the variable to which it points, and the two-headed arrows represent an association between the connected variables. One noteworthy feature of the path diagram is that it represents the random intercepts and random slopes as latent variables. They are latent because we do not directly observe the values of α_i and β_i , although we can estimate them. The time trend variable enters as the factor loadings in the SEM approach latent curve models.

When the data are continuous, it is common to assume that the repeated measures derive from a multivariate normal distribution. A maximum likelihood estimator (MLE) is available. In the SEM literature, methodologists have proposed corrections and adjustments to the MLE and significance tests that are asymptotically valid, even when the observed variables are not from a multinormal distribution. If the repeated measures are noncontinuous (e.g., dichotomous or ordinal), then alternative estimators must be used.

To illustrate the unconditional latent curve model, consider bimonthly data on logged total crime indexes in 616 communities in Pennsylvania for the first 8 months of 1995. These data come from the FBI Uniform Crime Reports and allow viewing the trajectory of crime over the course of the year. We estimate a mean value of 5.086 for the latent intercept, indicating that the mean logged crime rate at the first time point (January-February) is 5.086 for these communities.

The mean value of 0.105 for the latent slope indicates that the logged crime rate increases on average 0.105 units every 2 months over this time period. Both of these parameters are highly significant ($p < 0.01$). The significant intercept variance of 0.506 suggests that there is considerable variability in the level of crime in these communities at the beginning of the year, whereas the significant slope variance of 0.007 suggests variability in the change in crime over these 8 months.

CONDITIONAL MODEL

A second major type of latent curve model is the conditional model. Like the unconditional model, we can conceptualize the model as having Level 1 and Level 2 equations. Except in instances with time-varying covariates, the Level 1 equation is identical to that of the unconditional model (i.e., $y_{it} = \alpha_i + \beta_i \lambda_t + \varepsilon_{it}$). The difference occurs in the Level 2 equations. Instead of simply including the means of the intercepts and slopes and the deviations from the means, the Level 2 equations in the conditional latent curve model include variables that help determine the values of the intercepts and slopes. For instance, suppose that we hypothesize that two variables—say, x_{1i} and x_{2i} —affect the over-time trajectories via their impact on a case's intercept and slope. We can write the Level 2 equations as

$$\alpha_i = \mu_\alpha + \gamma_{\alpha x_1} x_{1i} + \gamma_{\alpha x_2} x_{2i} + \zeta_{\alpha i},$$

$$\beta_i = \mu_\beta + \gamma_{\beta x_1} x_{1i} + \gamma_{\beta x_2} x_{2i} + \zeta_{\beta i},$$

where μ_α and μ_β now are the intercepts of their respective equations, and the γ s are the regression coefficients that give the expected impact on the random intercept (α_i) or random slope (β_i) of a one-unit change in the x variable, controlling for the other variables in the equation. We still interpret the disturbances of each equation the same way and make the same assumptions with the addition that the disturbances are uncorrelated with the x s. Figure 2 is a path diagram of the conditional Level 1 and Level 2 equations, represented as an SEM.

An advantage of the conditional model is that a researcher can determine which variables influence the intercepts or slopes and hence better understand what affects the individual trajectories. The above Level 2 equations in the conditional model have only two x s, although in practice, researchers can use any number of x s.

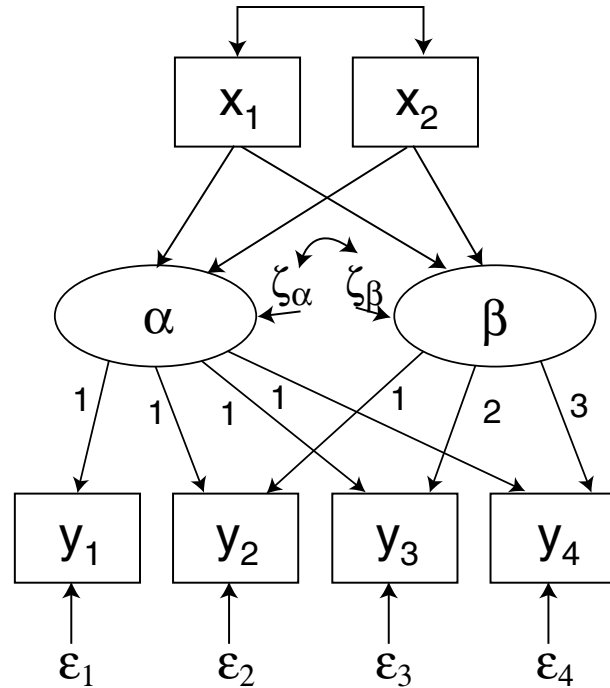


Figure 2 Conditional Latent Trajectory Model

An MLE for the conditional model is available when $\zeta_{\alpha i}$, $\zeta_{\beta i}$, and ε_{it} are distributed multivariate normal and the estimator is conditional on the value of the x s. As in the unconditional model, adjustments to the MLE permit significance testing when this distributional assumption does not hold.

Returning to the violent crime index, we include measures of population density and poverty rates for each community. The estimate of 0.02 ($p < 0.01$) for the effect of the poverty rate on the latent intercept suggests that a 1 percentage point difference in the poverty rate leads to an expected 0.02-unit difference in the level of logged crime at the first time point net of population density. However, the estimate of -0.002 for the effect of poverty on the latent slope indicates that a 1 percentage point increase in the poverty rate has an expected 0.002 decrease in the change in logged crime every 2 months net of population density. Population density has a positive effect on both the latent intercept and latent slope (although the latter narrowly misses significance), indicating that increasing population density increases the level of crime at the first time point and adds marginally to the slope of the trajectory in crime over this time period net of the poverty rate.

EXTENSIONS

There are many extensions to the unconditional and conditional latent curve models. One is to permit nonlinear trajectories in the outcome variable. This can be done by fitting polynomials in time, transforming the data, or, in the case of the SEM approach, freeing some of the factor loadings that correspond to the time trend so that these loadings are estimated rather than fixed to constants. In addition, autoregressive disturbances or autoregressive effects of the repeated measures can be added to the model. Researchers can incorporate time-varying predictors as a second latent curve or as a series of variables that directly affect the repeated measures that are part of a latent curve model. Direct maximum likelihood estimation or multiple imputation can handle missing data. Techniques to incorporate multiple indicators of a repeated latent variable also are available, as are procedures to include noncontinuous (e.g., categorical) outcome variables. In addition, recent work has investigated the detection of latent classes or groups in latent curve models. Furthermore, many measures of model fit are available, especially from the SEM literature. For further discussion of these models within the multilevel context, see Raudenbush and Bryk (2002); see also Bollen and Curran (in press) for a discussion from the SEM perspective. Software packages available to date for estimating latent growth curves include AMOS, EQS, LISREL, and MPLUS for the SEM approach, as well as HLM, MLn, and Proc Mixed in SAS for the hierarchical modeling approach. See the *Encyclopedia of Statistical Sciences* (Kotz & Johnson, 1983, pp. 539–542) and the *Encyclopedia of Biostatistics* (Armitage & Colton, 1998, pp. 3012–3015) for overviews of growth curve modeling in biostatistics and statistics.

—Kenneth A. Bollen, Sharon L. Christ,
and John R. Hipp

REFERENCES

- Armitage, P., & Colton, T. (Eds.). (1998). *Encyclopedia of biostatistics*. New York: John Wiley.
- Bollen, K. A., & Curran, P. J. (in press). *Latent curve models: A structural equation approach*. New York: John Wiley.
- Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica*, 25, 501–522.
- Kotz, S., & Johnson, N. (Eds.). (1983). *Encyclopedia of statistical sciences*. New York: John Wiley.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Wishart, J. (1938). Growth-rate determinations in nutritional studies with the Bacon pig, and their analysis. *Biometrika*, 30, 16–28.

GUTTMAN SCALING

Many phenomena in the social sciences are not directly measurable by a single item or variable. However, the researcher must still develop valid and reliable measures of these theoretical CONSTRUCTS in an attempt to measure the phenomenon under study. SCALING is a process whereby the researcher combines more than one item or variable in an effort to represent the phenomenon of interest. Many scaling MODELS are used in the social sciences.

One of the most prominent is Guttman scaling. Guttman scaling, also known as scalogram analysis and cumulative scaling, focuses on whether a set of items measures a single theoretical construct. It does so by ordering both items and subjects along an underlying cumulative dimension according to intensity. An example will help clarify the distinctive character of Guttman scaling. Assume that 10 appropriations proposals for the Department of Defense are being voted on in Congress—the differences among the proposals only being the amount of money that is being allocated for defense spending from \$100,000,000 to \$1,000,000,000 at hundred million-dollar increments. These proposals would form a (perfect) Guttman scale if one could predict how each member of Congress voted on each of the 10 proposals by knowing only the total number of proposals that each member of Congress supported. A scale score of 8, for example, would mean that the member supported the proposals from \$100,000,000 to \$800,000,000 but not the \$900,000,000 or \$1,000,000,000 proposals. Similarly, a score of 2 would mean that the member only supported the \$100,000,000 and \$200,000,000 proposals while opposing the other 8 proposals. It is in this sense that Guttman scaling orders both items (in this case, the 10 appropriations proposals) and subjects (in this case, members of Congress) along an underlying cumulative dimension according to intensity (in this case, the amount of money for the Department of Defense).

A perfect Guttman scale is rarely achieved; indeed, Guttman scaling anticipates that the perfect or ideal

model will be violated. It then becomes a question of the extent to which the empirical data deviate from the perfect Guttman model. Two principal methods are used to determine the degree of deviation from the perfect model: (a) minimization of error, proposed by Guttman (1944), and (b) deviation from perfect reproducibility, based on work by Edwards (1948). According to the minimization of error criterion, the number of errors is the least number of positive responses that must be changed to negative, or the least number of negative responses that must be changed to positive, for the observed responses to be transformed into an ideal response pattern. The method of deviation from perfect reproducibility begins with a perfect model and counts the number of responses that are inconsistent with that pattern. Error counting based on deviations from perfect reproducibility results in more errors than

the minimization of error technique and is a more accurate description of the data based on scalogram theory. For this reason, it is superior to the minimization of error method.

—Edward G. Carmines and James Woods

REFERENCES

- Carmines, E. G., & McIver, J. P. (1981). *Unidimensional scaling* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-024). Beverly Hills, CA: Sage.
- Edwards, A. (1948). On Guttman's scale analysis. *Educational and Psychological Measurement*, 8, 313-318.
- Guttman, L. L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139-150.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

H

HALO EFFECT

Also known as the *physical attractiveness stereotype* and the “*what is beautiful is good*” principle, halo effect, at the most specific level, refers to the habitual tendency of people to rate attractive individuals more favorably for their personality traits or characteristics than those who are less attractive. Halo effect is also widely used in a more general sense to describe the global impact of a likeable personality, or some specific desirable trait, in creating biased judgments of the target person on any dimension. Thus, feelings generally overcome cognitions when we appraise others.

This principle of judgmental bias caused by a single aspect of a person is grounded in the theory of central traits: Asch (1946) found that a man who had been introduced as a “warm” person was judged more positively on many aspects of his personality than when he had been described as “cold.”

Dion, Berscheid, and Walster (1972) asked students to judge the personal qualities of people shown to them in photos. People with good-looking faces, compared to those who were either unattractive or average-looking, were judged to be superior with regard to their personality traits, occupational status, marital competence, and social or professional happiness. This effect was not altered by varying the gender of the perceiver or the target person.

A reverse halo effect can also occur, in which people who possess admired traits are judged to be more attractive. A negative halo effect refers to negative bias produced by a person’s unattractive appearance. Halo

effect occurs reliably over a wide range of situations, involving various cultures, ages, and types of people. It affects judgments of social and intellectual competence more than ratings of integrity or adjustment (Eagly, Ashmore, Makhijani, & Longo, 1991).

Attempts to evaluate individuals’ job performance are often rendered invalid by halo effect: Ratings given by coworkers commonly reflect the attractiveness or likeability of the ratee rather than an objective measure of his or her actual competence. Halo effect also distorts judgments of the guilt of a defendant and the punishment he or she deserves. Raters may be given specific training in an attempt to reduce halo effect, very specific questions may be used, or control items may be included to estimate this bias.

The stereotype that attractive individuals possess better personal qualities is typically false in an experimental situation, where degree of beauty may be manipulated as a true INDEPENDENT VARIABLE and assigned randomly to individuals (e.g., by photos), but in real life, it may be true for certain traits. People who are beautiful, for example, may actually develop more confidence or social skills, creating a self-fulfilling prophecy.

—Lionel G. Standing

REFERENCES

- Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 41, 258–290.
- Dion, K. K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, 24, 285–290.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good but . . . A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, 110, 109–128.

HAWTHORNE EFFECT

The Hawthorne effect refers to a methodological artifact (see ARTIFACTS IN RESEARCH PROCESS) in behavioral FIELD EXPERIMENTS arising from the awareness of research participants that they are being studied. This awareness leads them to respond to the social conditions of the data collection process rather than to the experimental treatment the researcher intended to study. Somewhat akin to what has been called the “guinea pig effect” in laboratory research, the artifact reduces the INTERNAL VALIDITY of experiments. The term *Hawthorne effect* may also be used in a management context to refer to workers’ improved performance that is due to special attention received from their supervisor. Some researchers may confuse these usages, incorrectly interpreting improved performance due to a Hawthorne effect as a desirable outcome for their experiment.

Both effects were observed in a series of studies conducted in the 1920s and 1930s at the Hawthorne Works of the Western Electric Company (Roethlisberger & Dickson, 1939). Investigating the influence of various conditions on work performance, researchers were surprised that each new variable led to heightened levels of productivity, but were even more surprised when increased levels of productivity were still maintained after all improvements to working conditions had been removed. Researchers concluded that this unexpected result had been caused by incidental changes introduced in an attempt to create a controlled experiment.

Although there has been continuing controversy over the exact source of the artifact, and whether there was even evidence for it in the original experiments (Jones, 1992), the Hawthorne studies provided some of the first data suggesting the potential for social artifact to contaminate human research. The Hawthorne effect has been widely considered to be a potential contaminant of field experiments conducted in real-life settings, most notably in education, management, nursing, and social research in other medical settings. Researchers in a wide range of social sciences continue

to control for Hawthorne effects, or claim its potential for misinterpretation of their research results.

Rather than procedures preventing its occurrence, the extent of artifact in a study is typically addressed by the inclusion of CONTROL GROUPS to measure the difference in performance on the DEPENDENT VARIABLE between a no-treatment control group and a “Hawthorne” group. The conditions manipulated to create the Hawthorne control group are one of the variables typically thought to generate Hawthorne effects; the others are special attention, a novel (yet meaningless) task, and merely leading participants to feel that they are subjects of an experiment. The EFFECT SIZES associated with each of these Hawthorne control procedures in educational research (Adair, Sharpe, & Huynh, 1989) have been found by meta-analyses (see META-ANALYSIS) to be small and nonsignificant, suggesting that these control groups have been ineffective for assessing artifact.

In spite of this lack of evidence in support of any of its controls and the difficulties in defining the precise nature of the Hawthorne effect, the potential for artifact within field experiments remains. The significance of the Hawthorne studies was to suggest how easily research results could be unwittingly distorted simply by the social conditions within the experiment. In contemporary research, the term *Hawthorne effect* is often used in this broad sense to refer to nonspecific artifactual effects that may arise from participants’ reactions to experimentation.

—John G. Adair

REFERENCES

- Adair, J. G., Sharpe, D., & Huynh, C. L. (1989). Hawthorne control procedures in educational experiments: A reconsideration of their use and effectiveness. *Review of Educational Research*, 59, 215–228.
- Jones, S. R. (1992). Was there a Hawthorne effect? *American Journal of Sociology*, 98, 451–468.
- Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the worker*. Cambridge, MA: Harvard University Press.

HAZARD RATE

A hazard rate, also known as a *hazard function* or *hazard ratio*, is a concept that arises largely in the literatures on SURVIVAL ANALYSIS and EVENT HISTORY

ANALYSIS, and this entry focuses on its usage and estimation in these literatures. However, the estimator of the hazard ratio also plays a key role in the MULTIVARIATE modeling approach to correcting for sample SELECTION BIAS proposed by Heckman (1979).

The mathematical definition of a hazard rate starts with the assumption that a continuous RANDOM VARIABLE T defined over a range $t_{\min} \leq T \leq t_{\max}$ has a cumulative distribution function (CDF), $F(t)$, and a corresponding PROBABILITY DENSITY FUNCTION, $f(t) \equiv dF(t)/dt$. Frequently, the random variable T is assumed to be some measure of time. The complement of the cumulative probability function, called the survivor function, is defined as $S(t) \equiv 1 - F(t)$. Because it gives the probability that $T \geq t$ (i.e., “survives” from $t_{\min}$ to t), it is also known as the survivor probability. The hazard rate (or hazard ratio), $h(t)$, is then defined as

$$h(t) \equiv f(t)/S(t).$$

Speaking roughly and not rigorously, the hazard rate tells the chance or probability that the random variable T falls in an infinitesimally small range, $t \leq T < t + \Delta t$, divided by Δt , given that $T \geq t$, when Δt is allowed to shrink to 0. (Mathematically, one takes the limit as Δt approaches 0.)

Because both the probability density function and the survivor function cannot be negative, the hazard rate cannot be negative. The hazard rate is often regarded as analogous to a probability, and some authors even refer to a “discrete [time] hazard rate” (rather than to a probability) when T is taken to be a discrete random variable. However, for a continuous random variable T , the hazard rate can exceed 1, and, unlike a probability, it is not a unit-free number. Rather, its scale is the inverse of the scale of T . For example, in event history analysis, T might denote the age at which a person experiences some event (marriage, death) or the duration at which someone enters or leaves some status (starts or ends a job). In such instances, the hazard rate might be expressed in units “per year” or “per month.”

Because the survivor function equals 1 at $t_{\min}$, the hazard rate equals the probability density function at $t_{\min}$. In addition, because the survivor function is a monotonically nonincreasing function and equals 0 at $t_{\max}$ (unless T has a “defective” probability distribution), the hazard rate tends to become an increasingly larger multiple of the probability density function as t increases.

The LIFE TABLE or actuarial ESTIMATOR of the hazard rate in a specified discrete interval $[u, v]$ has a very long history of usage in demography and is still utilized in many practical applications. Its definition is

$$\widehat{h(t)} = \frac{d(u, v)}{(v - u)\{n(u) - \frac{1}{2}[d(u, v) + c(u, v)]\}}, \quad u \leq t < v$$

where $d(u, v)$ is the number of observations of t falling in the interval $[u, v]$; $n(u)$ is the number of observations that survived until u (i.e., for which $T \geq u$); and $c(u, v)$ is the number of observations censored in the interval $[u, v]$. (See CENSORING AND TRUNCATION for a definition and discussion of censored observations.) One-half appears in the divisor on the right-hand side of the equation because observations within the interval $[u, v]$ are assumed to occur uniformly over the interval.

Another widely used estimator is derived from the Nelson-Aalen estimator of the integrated hazard rate:

$$\widehat{h(t)} = \frac{d(u, v)}{(v - u)n(u)}, \quad u \leq t < v.$$

Cox and Oakes (1984) provide a clear, concise introduction to these and several other common nonparametric estimators of a hazard rate.

Multivariate models of hazard rates can be categorized in various ways. Some are fully parametric models; for example, the hazard rate may be postulated to be a particular function of EXPLANATORY VARIABLES and of the random variable T in a way that reduces to a simple Weibull or Gompertz model in t . Partially parametric models are widely used, in particular, the proportional hazard rate model proposed by Cox (1972). In these models, the hazard rate is assumed to be the product of a nuisance function, $q(t)$, and a parametric function of the explanatory variables, typically $\exp[\beta'x(t)]$.

Fully parametric models include not only proportional hazard rate models, but also nonproportional hazard rate models. For a discussion of these and other types of hazard rate models, along with some social scientific examples, see Tuma and Hannan (1984, Chaps. 3–8). Fully parametric models are usually estimated by MAXIMUM LIKELIHOOD; Cox’s proportional model is typically estimated by partial likelihood. In both fully and partially parametric models, the explanatory variables may be time invariant or time varying. The latter type demands either much more complete

temporal data on the explanatory variables or fairly restrictive assumptions about how the explanatory variables change when they are incompletely observed over time.

Exploratory analyses of data often suggest that a nonproportional hazard rate model would fit empirical data better than a proportional hazard rate model. Local hazard rate models, proposed by Wu and Tuma (1990), offer a way to model nonproportionality when the sample size is large. The article on the closely related concept of a TRANSITION RATE refers to yet another approach.

—Nancy Brandon Tuma

REFERENCES

- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187–220.
- Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. London and New York: Chapman and Hall.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- Tuma, N. B., & Hannan, M. T. (1984). *Social dynamics: Models and methods*. Orlando, FL: Academic Press.
- Wu, L. L., & Tuma, N. B. (1990). Local hazard models. In C. C. Clogg (Ed.), *Sociological methodology 1990* (pp. 141–180). Oxford, UK: Basil Blackwell.

HERMENEUTICS

Although hermeneutic literally means “making the obscure plain,” it is generally translated as “to interpret” or “to understand.” Hermeneutics emerged in 17th-century Germany as a method of biblical interpretation. Later, it was applied to other texts, particularly those that are obscure, alien, or symbolic. The aim is to uncover the meanings and intentions that are hidden in the text, including those of which even the author may not have been aware (known as *philological* hermeneutics). In the context of social research, hermeneutics is concerned with the interpretation of meaningful human action.

ORIGINS OF MODERN HERMENEUTICS

Friedrich Schleiermacher (1768–1834) provided the foundation for modern hermeneutics. Because he saw hermeneutics as a science for understanding any

utterance in language, hermeneutics moved from a concern with the analysis of texts from the past to the problem of how a person from one culture or historical period grasps the experiences of a person from another culture or period (known as general hermeneutics).

For Schleiermacher, understanding has two dimensions—grammatical and psychological. Grammatical interpretation corresponds to the understanding of language itself. Psychological interpretation attempts to recreate the act that produced the text and involves placing oneself within the mind of the author in order to know what was known and intended by this person as he or she wrote. Schleiermacher argued that the interpreter, as an outsider, is in a better position than the author to grasp and describe the “totality.”

From its background in scriptural and other textual interpretation, hermeneutics came to be seen as the core discipline for understanding all great expressions of human life, cultural and physical. The instigator of this transition, Wilhelm Dilthey (1833–1911), argued that the study of human conduct should be based on the method of understanding (*verstehen*) to grasp the subjective consciousness of the participants, whereas the study of natural phenomena should seek causal explanation (*erklären*). He rejected the methods of the natural sciences as being appropriate for the human sciences and addressed the question of how to achieve OBJECTIVITY in the human sciences. He set out to demonstrate the methods, approaches, and categories—applicable in all the human sciences—that would guarantee objectivity and VALIDITY. Whether he produced a satisfactory answer to the question is a matter of some debate.

In his early work, Dilthey hoped that the foundation of the human sciences would be based on descriptive psychology—an empirical account of consciousness devoid of concerns with causal explanation. He believed that psychology could provide a foundation for the social sciences in the same way that mathematics underlies the natural sciences. All human products, including culture, were seen to be derived from mental life. However, Dilthey came to realize the limits of this position and eventually moved from a focus on the mental life of individuals to understanding based on socially produced systems of meaning. He came to stress the role of social context and what he called “objective mind”—objectifications or externalizations of the human mind, or the “mind-created world,” that are sedimented in history, in what social scientists now call culture.

Dilthey now argued that phenomena must be situated in the larger wholes from which they derive their meaning; parts acquire significance from the whole, and the whole is given its meaning by the parts. The dual process of discovering taken-for-granted meanings from their externalized products, and understanding the products in terms of the meanings on which they are based, is known as the *hermeneutic circle*. The interpreter attempts to reconstruct the social and historical circumstances in which a text was produced and then to locate the text within these circumstances. Dilthey continued to assert the view that objective understanding must be the ultimate aim of the human sciences, even if the method is circular.

Dilthey insisted that the foundation for understanding human beings is in life itself, not in rational speculation or metaphysical theories. Life, or human experience, provides us with the concepts and categories we need to produce this understanding. He regarded the most fundamental form of human experience to be *lived experience*—first-hand, primordial, unreflective experience. According to Dilthey, the capacity of another person, or a professional observer, to understand human products is based on a belief that all human beings have something in common. However, he accepted the possibility that human expressions of one group may not be intelligible to members of another group; they may be so foreign that they cannot be understood. On the other hand, they may be so familiar that they do not require interpretation.

Martin Heidegger (1889–1976) was influenced by Dilthey and helped to lay the foundations for one branch of contemporary hermeneutics. Like Dilthey, he also wanted to establish a method that would reveal life in terms of itself. The central idea in Heidegger's work is that understanding is a mode of being rather than a mode of knowledge, an ONTOLOGICAL problem rather than an EPISTEMOLOGICAL one. It is not about how we establish knowledge; it is about how human beings exist in the world. For Heidegger, understanding is embedded in the fabric of social relationships, and interpretation is simply making this understanding explicit in language. Therefore, understanding is an achievement within reach of all human beings.

Heidegger recognized that there is no understanding of history outside of history. To assume that what is really there is self-evident is to fail to recognize the taken-for-granted presuppositions on which such assumed self-evidence rests. All understanding is

temporal; it is not possible for any human being to step outside history or his or her social world.

To recapitulate, early hermeneutics arose in order to overcome a lack of understanding of texts; the aim was to discover what the text means. Schleiermacher shifted the emphasis away from texts to an understanding of how members of one culture or historical period grasp the experiences of a member of another culture or historical period. He argued for a method of psychological interpretation, of re-experiencing the mental processes of the author of a text or speaker in a dialogue. This involves the use of the hermeneutic circle, or the process of grasping the unknown whole from the fragmented parts and using this to understand any part. Dilthey then shifted the emphasis to the establishment of a universal methodology for the human sciences, one that would be every bit as rigorous and objective as the methods of the natural sciences. He moved from psychological interpretation to the socially produced systems of meaning, from introspective psychology to sociological reflection, from the reconstruction of mental processes to the interpretation of externalized cultural products. Lived experience provides the concepts and categories for this understanding. For both Schleiermacher and Dilthey, because an interpreter's prejudices would inevitably distort his or her understanding, it is necessary to extricate oneself from entanglement in a sociohistorical context. Heidegger disagreed and regarded understanding as being fundamental to human existence and, therefore, the task of ordinary people. He argued that there is no understanding outside of history; human beings cannot step outside of their social world or the historical context in which they live.

TWO BRANCHES

The views of these early scholars reveal two polarized positions that divide on the claim of whether or not objective interpretation is possible. One, based on the work of Schleiermacher and Dilthey, looked to hermeneutics for general methodological principles of interpretation. It endeavored to establish an objective understanding of history and social life above and outside of human existence. The other, based on the work of Heidegger, regarded hermeneutics as a philosophical exploration of the nature and requirements for all understanding and regarded objectively valid interpretation as being impossible. He claimed that understanding is an integral part of everyday

human existence and is therefore the task of ordinary people, not of experts. Although these traditions have persisted, the latter has predominated and has been further developed by Hans-Georg Gadamer.

Gadamer (1989) was not particularly interested in the further development of hermeneutic methods for the social sciences, nor was he specifically concerned with the interpretation of texts. He focused on what is common to all modes of understanding by addressing three questions: How is “understanding” possible? What kinds of knowledge can understanding produce? What is the status of this knowledge?

For Gadamer, the key to understanding is grasping the “historical tradition,” or understanding and seeing the world at a particular time and in a particular place within which, for example, a text is written. It involves adopting an attitude that allows a text to speak to us while also recognizing that the tradition in which the text is located may have to be “discovered” from other sources.

When viewed in the context of disciplines such as anthropology and sociology, historical tradition can be translated as “culture” or “worldview,” and texts as records of conversations between social participants or with a researcher. Gadamer’s position requires us to look beyond what is said to what is being taken for granted while it is being said, to the everyday meaning of both the language used and the situations in which the conversations occur. The aim is to hear beyond the mere words.

Another important feature of Gadamer’s approach is the recognition that the process of understanding the products of other traditions or cultures cannot be detached from the culture in which the interpreter is located. He was critical of Dilthey’s attempt to produce an “objective” interpretation of human conduct. For Gadamer, a text or historical act must be approached from within the interpreter’s horizon of meaning, and this horizon will be broadened as it is fused with that of the act or text. The process of understanding involves a “fusion of horizons” in which the interpreter’s own horizon of meaning is altered as a result of the hermeneutical conversation with the old horizon through the dialectic of question and answer. The interpreter engages the text in dialogue, a process that transforms both the text and the interpreter. The interpreter is not trying to discover what the text really means by approaching it with an unprejudiced

open mind; he or she is not so much a knower as an experiencer for whom the other tradition opens itself.

For Gadamer, hermeneutics is about bridging the gap between our familiar world and the meaning that resides in an alien world. Collision with another horizon can make the interpreter aware of his or her own deep-seated assumptions and his or her own prejudices or horizon of meaning, of which he or she may have remained unaware; taken-for-granted assumptions can be brought to critical self-consciousness, and genuine understanding can become possible.

Gadamer argued that understanding what other people say is not a matter of “getting inside” their heads and reliving their experiences. Because language is the universal medium of understanding, understanding is about the translation of languages. However, every translation is, at the same time, an interpretation. Every conversation presupposes that the two speakers speak the same language and understand what the other says. But the hermeneutic conversation is usually between different languages, ranging from what we normally regard as “foreign” languages, to differences due to changes in a language over time or to variations in dialect. The hermeneutic task is not the correct mastery of another language but the mediation between different languages. Thus, understanding as the fusion of horizons occurs through language; language allows the mediation, the interpenetration and transformation of past and present. Whether it be a conversation between two people or between an interpreter and a text, a common language must be created. Like Heidegger, Gadamer was interested in ontology, not epistemology, with establishing what all ways of understanding have in common rather than being concerned with problems of method.

Hermeneutics has provided the foundation for some branches of interpretivism, such as the work of Max Weber and Alfred Schütz, and for the logic of abduction.

—Norman Blaikie

REFERENCES

- Bauman, Z. (1978). *Hermeneutics and social science*. London: Hutchinson.
- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.

- Gadamer, H-G. (1989). *Truth and method* (rev. 2nd ed.). New York: Crossroad.
- Outhwaite, W. (1975). *Understanding social life: The method called verstehen*. London: Allen & Unwin.
- Palmer, R. E. (1969). *Hermeneutics: Interpretation theory of Schleiermacher, Dilthey, Heidegger, and Gadamer*. Evanston, IL: Northwestern University Press.

HETEROGENEITY

Statistics is the study of heterogeneity, so this is a broad topic. Statistical models typically distinguish between observed and unobserved sources of heterogeneity, capturing observed sources in terms of predictors or explanatory variables and relegating unobserved sources to the error term, typically assumed independent of observed covariates. Heterogeneity also arises when individual observations follow different models, leading to a distinction between subject-specific and population-average models. Because these ideas first became popular in demographic applications, we describe them in terms of survival or time-to-event data. The seminal paper here is Vaupel, Manton, and Stallard (1979).

Consider a homogeneous population where the risk that an event such as death will occur at time (or age) t , given that it has not yet occurred, is $\lambda(t)$, both for individuals and for the population as a whole. Consider next a heterogeneous population, and assume that the risk for individual i is $\theta_i \lambda(t)$, where $\theta_i > 0$ is an unobservable representing an individual's frailty, assumed to have a DISTRIBUTION with a mean of 1. It turns out that in such a heterogeneous population, the average HAZARD at age t is not $\lambda(t)$ but $E(\theta_i | T_i > t) \lambda(t)$. The multiplier is the average frailty of survivors to age t , is always less than 1, and declines with t . This implies that even if the risk were CONSTANT for each individual, the average risk in the POPULATION would decline over time as the weaker die first and leave exposed to risk a group increasingly selected for its robustness.

The fact that the risk one observes in a heterogeneous population is not the same as the risk faced by each individual can lead to interesting paradoxes. It has also led to attempts by analysts to control for unobserved heterogeneity by introducing frailty effects in their models, often in addition to observed covariates. ESTIMATION usually proceeds by specifying parametric forms for the hazard and the distribution of

frailty. Concerned that the estimates might be heavily dependent on ASSUMPTIONS about the distribution of the unobservable, Heckman and Singer (1984) proposed a nonparametric ESTIMATOR of the distribution of θ_i . But estimates can also be sensitive to the choice of parametric form for the hazard. Unfortunately, one cannot relax both assumptions, because then the model is not identified. Specifically, unobserved heterogeneity is confounded with negative duration dependence in models without covariates, and with the usual assumption of proportionality of hazards in models with covariates; see Rodríguez (1994) for details.

The situation is much better if one has repeated events per individual and frailty is persistent across events, or if one has hierarchical data such as children nested within mothers and frailty is shared by members of a family. In both situations, frailty models are fully identified. For a discussion of multivariate survival, see Hougaard (2000). These models can be viewed as special cases of the models used in MULTILEVEL ANALYSIS. The classical frailty framework corresponds to a random-intercept model, where the mean level of response varies across groups. One can also consider random-slope models, where the effect of a covariate on the outcome, such as a treatment effect, varies across groups. Alternatively, the assumption that frailty is independent of observed covariates can be relaxed by treating θ_i as a fixed rather than a random effect (see FIXED-EFFECTS MODEL).

—Germán Rodríguez

REFERENCES

- Heckman, J. J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52, 271–319.
- Hougaard, P. (2000). *Analysis of multivariate survival data*. New York: Springer-Verlag.
- Rodríguez, G. (1994). Statistical issues in the analysis of reproductive histories using hazard models. In K. L. Campbell & J. W. Wood (Eds.), *Human reproductive ecology: Interaction of environment, fertility and behavior* (pp. 266–279). New York: New York Academy of Sciences.
- Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439–454.

HETEROSCEDASTICITY. See
HETEROSKEDASTICITY

HETEROSKEDASTICITY

An assumption necessary to prove that the ORDINARY LEAST SQUARES (OLS) estimator is the BEST LINEAR UNBIASED ESTIMATOR (BLUE) is that the variance of the error is constant for all observations,

or HOMOSKEDASTIC. Heteroskedasticity arises when this assumption is violated because the error variance is nonconstant. When this occurs, OLS parameter estimates will be unbiased, but the variances of the estimated parameters will not be efficient. A classic example is the regression of income on savings (Figure 1 provides a hypothetical visual example). We expect

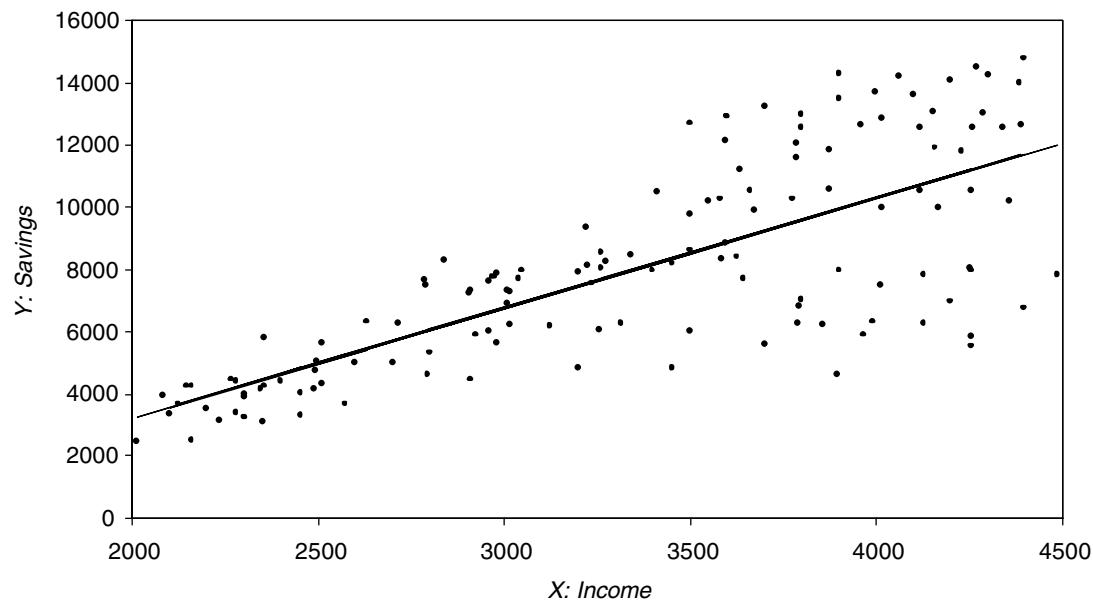


Figure 1 Heteroskedastic Regression

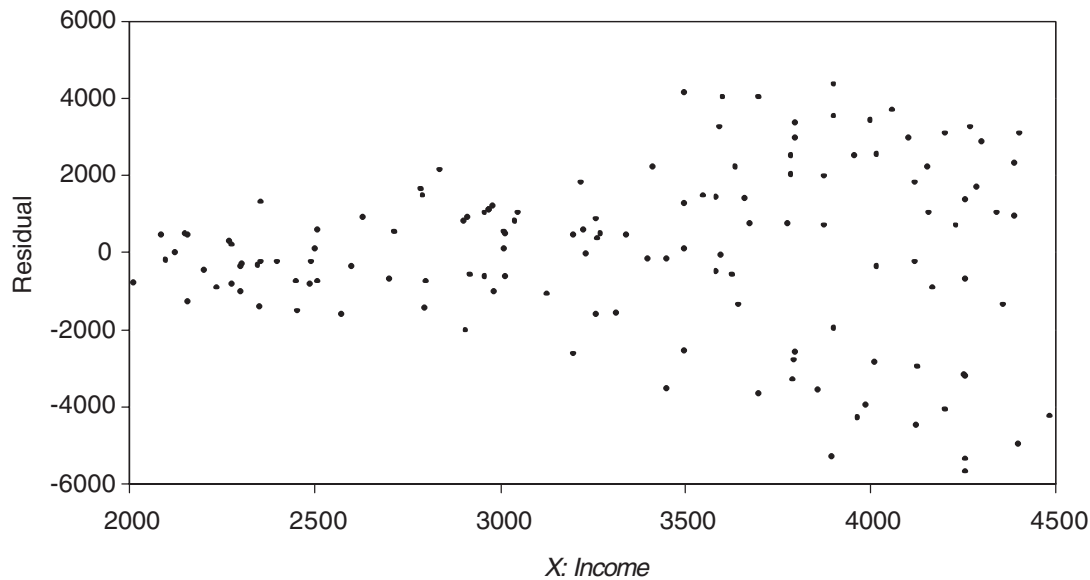


Figure 2 Visualizing Residuals

savings to increase as the level of income rises. This regression is heteroskedastic because individuals with low incomes have little money to save, and thus all have a relatively small amount of savings and small errors (they are all located near the regression line). However, individuals with high incomes have more to save, but some choose to use their surplus income in other ways (for instance, on leisure), so their level of savings varies more and the errors are larger.

A number of methods to detect heteroskedasticity exist. The simplest is a visual inspection of the residuals $\hat{u}_i = (y_i - \hat{y}_i)$ plotted against each independent variable. If the absolute magnitudes of the residuals seem to be constant across values of all independent variables, heteroskedasticity is likely not a problem. If the dispersion of errors around the regression line changes across any independent variable (as Figure 2 shows it does in our hypothetical example), heteroskedasticity may be present. In any event, more formal tests for heteroskedasticity should be performed.

The Goldfeld-Quandt and Breusch-Pagan tests are among the options. In the former, the observations are ordered according to their magnitude on the offending variable, some central observations are omitted, and separate regressions are performed on the two remaining groups of observations. The ratio of their sum of squared residuals is then tested for heteroskedasticity. The latter test is more general, in that it does not require knowledge of the offending variable.

If the tests reveal the presence of heteroskedasticity, a useful first step is to again scrutinize the residuals to see if they suggest any omitted variables. It is possible that all over-predictions have a trait in common that could be controlled for in the model. In the example, a measure of the inherent value individuals place on saving money (or on leisure) could alleviate the problem. The high income-low savings individuals may value saving less (or leisure more) than the average individual. If no such change in model specification is apparent, it may be necessary to transform the variables and conduct a GENERALIZED LEAST SQUARES (GLS) estimation. The GLS procedure to correct for heteroskedasticity weights the observations on the offending independent variable by the inverse of their errors and is often referred to as WEIGHTED LEAST SQUARES (WLS). If the transformation is done, and no other regression assumptions are violated, transformed estimates are BLUE.

—Kevin J. Sweeney

REFERENCES

- Berry, W. D. (1993). *Understanding regression assumptions*. Newbury Park, CA: Sage.
 Gujarati, D.N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.
 Kennedy, P. (1992). *A guide to econometrics* (3rd ed.). Cambridge: MIT Press.

HEURISTIC

A tool, device, or rule to facilitate the understanding of a concept or method. For example, to help students learn about SIGNIFICANCE TESTING, they may be offered the following heuristic: In general, a coefficient with a t ratio of 2.0 or more in absolute value will be significant at the .05 level. This heuristic is useful for sorting through t ratios that may have been generated by many REGRESSION runs.

—Michael S. Lewis-Beck

HEURISTIC INQUIRY

Heuristic inquiry is a form of phenomenological inquiry that brings to the fore the personal experience and insights of the researcher. With regard to some phenomenon of interest, the inquirer asks, “What is *my* experience of this phenomenon and the essential experience of others who also experience this phenomenon intensely?” (Patton, 2002).

Humanist psychologist Clark Moustakas (1990) named heuristic inquiry when he was searching for a word that would encompass the processes he believed to be essential in investigating human experience. “Heuristic” comes from the Greek word *heuriskein*, meaning “to discover” or “to find.” It connotes a process of internal search through which one discovers the nature and meaning of experience and develops methods and procedures for further investigation and analysis. “The self of the researcher is present throughout the process and, while understanding the phenomenon with increasing depth, the researcher also experiences growing self-awareness and self-knowledge. Heuristic processes incorporate creative self-processes and self-discoveries” (Moustakas, 1990, p. 9).

There are two focusing elements of heuristic inquiry within the larger framework of phenomenology: (a) The researcher *must* have personal experience with and intense interest in the phenomenon under study, and (b) others (coresearchers) who share *an intensity* of experience with the phenomenon participate in the inquiry. Douglass and Moustakas (1985) have emphasized that “heuristics is concerned with meanings, not measurements; with essence, not appearance; with quality, not quantity; with experience, not behavior” (p. 42).

The particular contribution of heuristic inquiry is the extent to which it legitimizes and places at the fore the personal experiences, reflections, and insights of the researcher. The researcher comes to understand the essence of the phenomenon through shared reflection and inquiry with coresearchers as they also intensively experience and reflect on the phenomenon in question. A sense of connectedness develops between researcher and research participants in their mutual efforts to elucidate the nature, meaning, and essence of a significant human experience.

The rigor of heuristic inquiry comes from systematic observation of and dialogues with self and others, as well as depth interviewing of coresearchers.

Heuristic inquiry is derived from but different from PHENOMENOLOGY in four major ways:

1. Heuristics emphasizes connectedness and relationship, whereas phenomenology encourages more detachment in analyzing experience.
2. Heuristics seeks essential meanings through portrayal of personal significance, whereas phenomenology emphasizes definitive descriptions of the structures of experience.
3. Heuristics includes the researcher’s intuition and tacit understandings, whereas phenomenology distills the structures of experience.
4. “In heuristics the research participants remain visible. . . . Phenomenology ends with the essence of experience; heuristics retains the essence of the person in experience” (Douglass & Moustakas, 1985, p. 43).

Systematic steps in the heuristic inquiry process lead to the exposition of experiential essence through immersion, incubation, illumination, explication, and creative synthesis (Moustakas, 1990).

—Michael Quinn Patton

REFERENCES

- Douglass, B., & Moustakas, C. (1985). Heuristic inquiry: The internal search to know. *Journal of Humanistic Psychology*, 25, 39–55.
- Moustakas, C. (1990). *Heuristic research: Design, methodology, and applications*. Newbury Park, CA: Sage.
- Moustakas, C. (1994). *Phenomenological research methods*. Thousand Oaks, CA: Sage.
- Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks, CA: Sage.

HIERARCHICAL (NON)LINEAR MODEL

The hierarchical linear model (HLM) is basic to the main procedures of MULTILEVEL ANALYSIS. It is an extension of the GENERAL LINEAR MODEL to data sets with a hierarchical nesting structure, where each level of this hierarchy is a source of unexplained variation. The general linear model can be formulated as

$$Y_i = b_0 + b_1 X_{1i} + \cdots + b_p X_{pi} + E_i,$$

where i indicates the case; Y is the dependent variable; X_1 to X_p are the independent (or explanatory) variables; E is the unexplained part, usually called the residual or error term; b_0 is called the intercept; and b_1 to b_p are the regression coefficients. Sometimes, the constant term b_0 is omitted from the model. The standard assumption is that the residuals E_i are independent, normally distributed, random variables with expected value 0 and a common variance, and the values X_{hi} are fixed and known quantities.

Examples of nesting structures are pupils nested in classrooms (a two-level structure), pupils nested in classrooms in schools (three levels), pupils nested in classrooms in schools in countries (four levels), and so on. Elements of such nesting structure are called units (e.g., pupils in this example are level-one units, classrooms are level-two units, etc.). For a two-level nesting structure, it is customary to use a double indexation so that, in this example, Y_{ij} indicates the dependent variable for pupil i in classroom j . The hierarchical linear model for two levels of nesting can be expressed as

$$Y_{ij} = b_0 + b_1 X_{1ij} + \cdots + b_p X_{pij} + U_{0j} + U_{1j} Z_{1ij} + \cdots + U_{qj} Z_{qij} + E_{ij}.$$

The first part, $b_0 + b_1 X_{1ij} + \dots + b_p X_{pij}$, is called the fixed part and has the same structure and interpretation as the first part of the general linear model; the double indexation of the explanatory variables is only a matter of notation. The second part, called the random part, is different. Unexplained variation between the level-two units is expressed by $U_{0j} + U_{1j} Z_{1ij} + \dots + U_{qj} Z_{qij}$, where Z_1 to Z_q are explanatory variables (usually, but not necessarily, a subset of X_1 to X_p); U_{0j} is called the random intercept; and U_{1j} to U_{qj} are called random slopes (it is allowed that there are no random slopes, i.e., $q = 0$). The variables U_{0j} to U_{qj} are residuals for level two, varying between the level-two units, and E_{ij} is a residual for level one. The standard assumption is that residuals for different levels are independent, residuals for different units at the same level are independent, and all residuals are normally distributed. It is possible to extend the model with random slopes at level one, in addition to the residual E_{ij} ; this is a way of formally representing HETEROSKEDASTICITY of the level-one residuals (see Snijders & Bosker, 1999, Chap. 8).

This model can be extended to more than two levels of nesting. For three nesting levels, with units being indicated by i , j , and k , the formula reads

$$Y_{ijk} = b_0 + b_1 X_{1ijk} + \dots + b_p X_{pijk} + V_{0k} + V_{1k} W_{1ijk} + \dots + V_{rk} W_{rijk} + U_{0jk} + U_{1jk} Z_{1ijk} + \dots + U_{qjk} Z_{qijk} + E_{ijk}.$$

The W_{hijk} now are also explanatory variables, with random slopes V_{hk} varying between the level-three units k . The rest of the model is similar to the two-level model. The hierarchical linear model also can be extended to sources of unexplained variation that are not nested but crossed factors (see the entries on MULTILEVEL ANALYSIS and MIXED-EFFECTS MODEL).

HIERARCHICAL NONLINEAR MODEL

The hierarchical nonlinear model, also called hierarchical generalized linear model or generalized linear mixed model, is an analogous extension to the GENERALIZED LINEAR MODEL. It is defined by including random residuals for the higher-level units in the linear predictor of the generalized linear model. Thus, the formulas given above for Y_{ij} and Y_{ijk} , but without the level-one residual E_{ij} or E_{ijk} , take the place of the elements of

the vector defined as the linear predictor $\eta = X\beta$ in the entry on generalized linear models.

—Tom A. B. Snijders

REFERENCES

- Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Hodder Arnold.
- Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

HIERARCHY OF CREDIBILITY

A term used by Howard Becker (1967) to refer to a differential capacity for the views of groups in society to be regarded as plausible. The term relates to the tendency for many spheres of society to have a clear hierarchy: police/criminal, manager/worker, teacher/student. It has been suggested that when social scientists take the worldviews of those groups that form the subordinate groups in these pairings, they are more likely to be accused of bias. This arises because of the hierarchy of credibility, whereby those in the superordinate positions are perceived by society at large as having the right and the resources for defining the way things are.

—Alan Bryman

REFERENCE

- Becker, H. S. (1967). Whose side are we on? *Social Problems*, 14, 239–247.

HIGHER-ORDER. See ORDER

HISTOGRAM

Histograms, BAR GRAPHS, and PIE CHARTS are used to display results in a simple, visual way. A bar chart would be used for categorical (nominal) variables, with each bar of the chart displaying the value of some variable. A histogram performs the

same function for grouped data with an underlying CONTINUOUS VARIABLE (an ordinal, interval, or ratio distribution). The bars touch to reflect the continuous nature of the underlying distribution. A histogram or bar chart would be used in preference to a pie chart for showing trends across a variable (e.g., over time or by geographical locality). Pie charts are better for showing the percentage composition of a population.

Figure 1 shows the number of moves (changes of householder) recorded against properties where there has been a change of householder in the past year, in the center of a town in northeastern England that has been in decline since the collapse of its major industries but is now undergoing considerable regeneration. The figure shows nearly 150 properties that have had two changes of householder during the year and nontrivial numbers where ownership has been even more volatile. (The data for this figure are taken from Hobson-West & Sapsford, 2001.) As well as being descriptively useful, this kind of graph helps to determine the overall shape of the variable—the extent

of departure from normality—which in this case is very substantial.

—Roger Sapsford

REFERENCE

Hobson-West, P., & Sapsford, R. (2001). *The Middlesbrough Town Centre Study: Final report*. Middlesbrough, UK: University of Teesside, School of Social Sciences.

HISTORICAL METHODS

Historical methods refers to the use of primary historical data to answer a question. Because the nature of the data depends on the question being asked, data may include demographic records, such as birth and death certificates; newspapers articles; letters and diaries; government records; or even architectural drawings.

The use of historical data poses several broad questions:

1. Are the data appropriate to the theoretical question being posed?
2. How were these data originally collected, or what meanings were embedded in them at the time of collection?
3. How should these data be interpreted, or what meanings do these data hold now?

THEORETICAL QUESTIONS

One way for social scientists to pose questions about contemporary issues is to compare aspects of contemporary society with those of past societies. As C. Wright Mills put it, to understand social life today, one must ask such questions as, “In what kind of society do I live?” and “Where does this society stand in the grand march of human history?” These questions necessarily imply a theory, a set of related propositions that orients the researcher as to where to search for answers. Indeed, to generalize, *any historical question implies a theory*, for without a theory, the social scientist is haphazardly collecting a conglomeration of shapeless facts.

I stress the centrality of an explicit theory to historical method to highlight a key distinction between

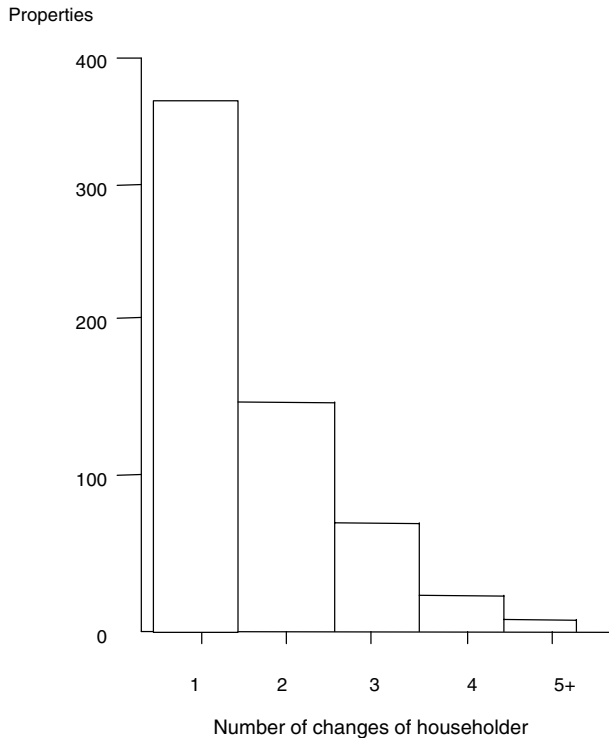


Figure 1 Number of Changes of Householder in Middlesbrough During 2001

historical methods used by historians and those used by social scientists. Although historians use theories, their theoretical assumptions are not always explicitly developed. Rather, their theoretical emphases are often contained in the professional classification in which they are placed—say, “military historian,” “diplomatic historian,” or “social historian”—because these classifications implicitly claim that the activities of a specific group (and hence a particular kind of power or lack of power) is central to the analysis of epochs. Within this profession, as within all academic professions, the centrality of any one category has varied over time. In the late 20th century, the advent of social history (including such specialties as women’s history and African American history) was an announcement that understanding people who had been historically powerless, even invisible, is vital to grasping the nature of our own time.

APPROPRIATE DATA

Such shifts in categorical emphasis also imply new forms of data. Thus, news reports have less utility to understand the lives of peasants, factory workers, slaves, or maids than they do to write the biography of, say, a military officer—although news stories may tell us how factory workers or slaves were seen by those who had more power than they. Similarly, one can *hope* to understand the differential reaction of European Americans and African Americans to urban riots by comparing mass media and African American media. But unless such reports are placed in a theoretical and comparative context, the researcher can only have *hope*.

Unfortunately, many social scientists embarking on historical inquiries have not been trained in history. In part, this deficiency means that when reading historical accounts, the social scientist may not know either the specific questions those accounts are debating or even the implicit theories that historians are addressing. Without a context, it is difficult to know what to remember, or even what to write on one’s ubiquitous note cards or computer files. Similarly, without some appreciation of the lives of people in a specific group living in a specific time and place, one cannot know the meaning or significance of a specific piece of data.

Not having been trained in history also poses another problem: A social scientist may not know the location of archives pertinent to a research problem. Taking a course in historiography, consulting an

accomplished historian, checking *Historical Abstracts*, and searching the Internet are all ways to learn what data others have used and thus what types of data may be available.

AN EXAMPLE

Some social scientists work inductively rather than deductively. If a social scientist is proceeding inductively, he or she must know the context. Let me give a seemingly insignificant example—constructing a family tree (an atheoretic enterprise). To delineate a family tree, I recalled my parents’, uncles’, and aunts’ stories about their childhood. Then, I visited my family’s graveyard, pen and paper in hand. I did so because I knew to expect that under each person’s English name, I would find his or her name in Hebrew, and that name would follow the expected form “first name middle name, son or daughter of first name middle name,” as in “Chava Perl bat [daughter] of Yisroel Avram.” By locating the gravestone of Yisroel Avram, I could then learn the name of his father and so on. One stone did not give the name of the deceased father. Rather, in Yiddish (the wrong language), that stone announced “daughter of a good man.” An elderly relative had once told me a family secret about the woman whose stone contained Yiddish: She had married her uncle. Examining the placement of graves (whose conventions I knew), I could infer the name of her uncle and also that of her mother. Without having background information, I could not have constructed the family tree. That family tree is simply data waiting for a theoretical frame.

OTHER CONTEXTUAL ISSUES

Contextual questions may be difficult in other ways as well. Reading histories, one wants to know what the historians are debating. (If the historians did not have a point, they would not have bothered to write.) Other nonfictions raise other questions: Reading 19th-century women’s diaries requires asking, Was this skill taught in school? Did a teacher read this diary as part of a class assignment? What kind of self-censorship (or external censorship) was involved in the production of any written document? Context—what is said and what is missing—governs meaning in historical data as in ethnographic studies of contemporary life. As is true of any qualitative method, context guides the recognition of theoretically pertinent

data, and theory directs the understanding of pertinent contexts.

INTERPRETATION

Theory is particularly pertinent to the interpretation of data, because all theories imply an EPISTEMOLOGICAL stance toward the very existence of data. There are at least two broad ways of viewing interpretative acts. I call them *reproduction* and *representation*. By reproduction, I mean the view that historical accounts accurately capture the essence of a specific time and place. I include here empiricist epistemologies, such as those that emphasize “the grand march of history.” Representation refers to the great variety of POST-MODERNIST views that identify both historical and social scientific writings as “texts,” which, as is the case of the data they analyze, are political documents that purposively or inadvertently take sides in struggles for power. However, more discussion about the differences between these broad schools of thought departs from a discussion of historical method.

—Gaye Tuchman

REFERENCES

- Himmelfarb, G. (1987). *The new history and the old*. Cambridge, MA: Harvard University Press.
- Marcus, G., & Fischer, M. (1986). *Anthropology as cultural critique*. Chicago: University of Chicago Press.
- Scott, J. W. (1988). *Gender and the politics of history*. New York: Columbia University Press.
- Tuchman, G. (1994). Historical social science: Methodologies, methods, and meanings. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 306–323). Thousand Oaks, CA: Sage.

HOLDING CONSTANT. See CONTROL

HOMOSCEDASTICITY. See HOMOSKEDASTICITY

HOMOSKEDASTICITY

For ORDINARY LEAST SQUARES (OLS) assumptions to be met, the ERRORS must exhibit constant variance. Such uniform variance is called homoskedasticity. If this assumption is not met, that is, if the errors exhibit nonuniform variance, the opposing condition of HETEROSKEDASTICITY obtains, and OLS is no longer the BEST LINEAR UNBIASED ESTIMATOR (BLUE) of model PARAMETERS. Specifically, if the errors are heteroskedastic, the estimated model parameters will remain UNBIASED, but they will not be minimum variance estimates.

To understand the concept of uniform error variance, it is helpful to consider the relationship between two variables, X and Y , and the conditional means and variances of Y with respect to X in a regression. If DEPENDENT VARIABLE Y is related to INDEPENDENT VARIABLE X , then mean values of Y will vary across different values of X . This is the conditional mean of Y given X , written $(Y|X)$. It is known that the regression line of Y on X passes directly through these conditional means. We may then consider the variation in Y at different conditional means. This is known as conditional variation, denoted $\text{var}[Y|X]$. This conditional variation is also referred to as the skedastic function. When homoskedastic, conditional variation will not change across values of X ; rather, it will be uniform. This uniform variation is denoted σ^2 (Figure 1 illustrates). The conditional mean of Y given particular values of X ($Y|X$) increases with X . But the dispersion of errors around those conditional means remains constant throughout.

If the errors are homoskedastic and exhibit no AUTOCORRELATION with each other, they are said to be “spherical.” With spherical DISTURBANCE TERMS, the EXPECTED VALUE of the error variance-covariance matrix, $E(UU')$, reduces to $\sigma^2\mathbf{I}$, and the OLS formulas for model parameters and their STANDARD ERRORS are obtained readily (see SPHERICITY ASSUMPTION). If the disturbances are nonspherical, then OLS becomes problematic and a GENERALIZED LEAST SQUARES (GLS) solution is required.

Because the errors are simply the difference between predicted and observed values of the dependent variable Y , they are themselves expressed in units of Y . Nevertheless, there is no reason to expect

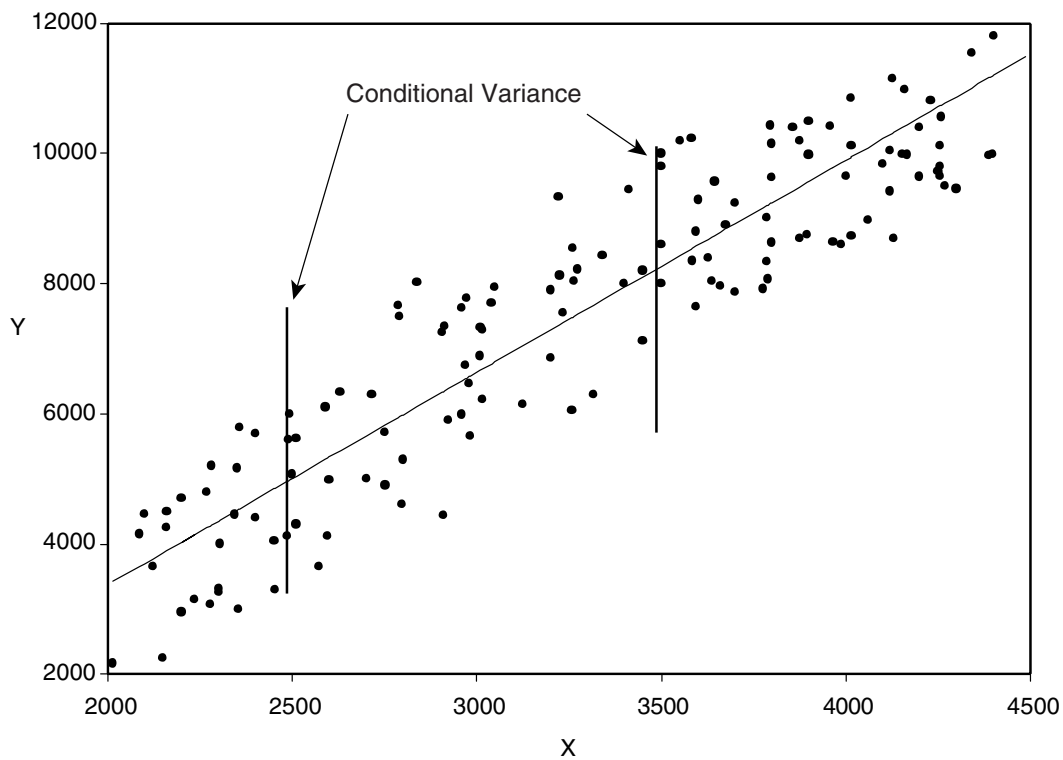


Figure 1 Homoskedastic Errors in Regression

that overall variation in Y , $\text{var}(Y)$, will be equal to $\text{var}(Y|X)$, even when that conditional variation is uniform (σ^2). It is expected that our model has explained some of the overall variation in Y , and it is only the conditional variation of the residuals that needs to be uniform (i.e., homoskedastic) for this OLS assumption to be met.

—Brian M. Pollins

REFERENCES

- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Gujarati, D. N. (2003). *Basic econometrics* (4th ed.). Boston: McGraw-Hill.
- Kennedy, P. (1998). *A guide to econometrics* (4th ed.). Cambridge: MIT Press.

HUMANISM AND HUMANISTIC RESEARCH

Humanistic research is research that gives prime place to human beings, human meaning, and human

actions in research. It usually also works with a strong ethical framework that both respects human beings and seeks to improve the state of humankind in a global context. It allies itself more with the humanities than with science, which it often critiques.

There is, of course, a long history of diverse forms of humanism in world history. The Greek Sophists and Socrates “called philosophy down from heaven to earth” (as Cicero put it). Early humanists of the pre-Renaissance movement, such as Erasmus, did not find their humanism incompatible with religion, but they did sense the abilities of human beings to play an active role in creating and controlling their world. The Italian Renaissance itself was a period in which the study of the works of man—and not simply God—became more and more significant. French philosophers, such as Voltaire, gave humanism a strongly secular base. Alfred McLung Lee (1978), a champion of a humanistic sociology, sees humanism “in a wide range of religious, political and academic movements” (p. 44) and links it to fields as diverse as communism, democracy, egalitarianism, populism, NATURALISM, POSITIVISM, pragmatism, relativism, science, and supernaturalism, including versions of ancient

paganisms, Hinduism, Buddhism, Judaism, Roman Catholicism, Protestantism and Mohammedanism (pp. 44–45). It does seem to be one of those words that can mean all things to all people.

In the social sciences, there are clear humanistic strands in psychology (the works of Abraham Maslow, Jerome Bruner, and Robert Coles); anthropology (Gregory Bateson, Margaret Mead, Robert Redfield); and sociology (C. Wright Mills, Peter Berger, the symbolic interactionists). Humanistic psychologists, for example, criticize much of psychology as dehumanizing and trivial, providing elegant experiments but failing to deal with the major ethical crises of our time or the ways in which living, whole human beings deal with them. Within sociology, humanistic research is frequently linked to research that examines the active construction of meanings in people's lives.

As with William James's *Varieties of Religious Experience* or Margaret Mead's *Sex and Temperament in Three Primitive Societies*, humanistic social science usually recognizes the varieties of these meanings. Often employing field research, life stories, and qualitative research—as well as using wider sources such as literature, biographies, photographs, and film—it aims to get a close and intimate familiarity with life as it is lived. Humanistic research tends to shun abstractions and adopt a more pragmatic, down-to-earth approach. Although it usually places great faith in human reason, it also recognizes the role of emotions and feelings, and it puts great emphasis upon the ethical choices people make in creating the good world and the good life.

Humanistic research is attacked from many sides. Scientists have tended to deride its lack of objectivity, determinism, and technical competence. Nihilist philosophers like Nietzsche and Kierkegaard shun its overly romantic view of the human being and dismiss its ethical systems. Many religions are critical because it fails to see humans as worthless and only saved by salvation or divine grace. Most recently, multiculturalists and postmodernists have been concerned that it is often limited to a very specific Western and liberal version of what it is to be human. Criticisms notwithstanding, it acts as a subversive tradition to major scientific tendencies within the social sciences.

—Ken Plummer

See also AUTOBIOGRAPHY, INTERPRETATIVE BIOGRAPHY, LIFE HISTORY INTERVIEW, LIFE HISTORY RESEARCH, ORAL HISTORY, QUALITATIVE RESEARCH, REFLEXIVITY

REFERENCES

- Lee, A. M. (1978). *Sociology for whom*. New York: Oxford University Press.
- Radest, H. B. (1996). *Humanism with a human face: Intimacy and the enlightenment*. London: Praeger.
- Rengger, J. (1996). *Retreat from the modern: Humanism, post-modernism and the flight from modernist culture*. Exeter, UK: Bowerdean.

HUMANISTIC COEFFICIENT

In contrast with a social science that sees the world in terms of “facts” and “things,” there is a persistent strand that is concerned with *meaningful, intersubjective social actions and values*. Max Weber's concern with VERSTEHEN is one of the most celebrated of such approaches. The work of Alfred Schutz on PHENOMENOLOGY is another.

During the 1920s and 1930s, another key figure in clarifying this position was Florian Znaniecki (although he is largely neglected today). Both in his methodological note to *The Polish Peasant* (with W. I. Thomas) and in his *The Method of Sociology* (1934), he presents a strong concern with the neo-Kantian distinction between two systems, natural and cultural. Natural systems are given objectively in nature and have an independent existence. They are separate from the experience and activities of people. Cultural systems are intrinsically bound up with the conscious experiences of human agents in interaction with each other.

For Znaniecki (1934), “natural systems are objectively given to the scientist as if they existed absolute independently of the experience and activity of men” (p. 136). They include such things as the geological composition of the earth, the chemical compound, or the magnetic field. They can exist without any participation of human consciousness. “They are,” as he says, “bound together by forces which have nothing to do with human activity” (p. 136). In stark contrast, cultural systems deal with the human experiences of conscious and active historical subjects. This cultural system is believed to be real by these historical subjects experiencing it; but it is not real in the same way as the natural system is. The object of study is always linked to somebody's human meanings. Znaniecki (1934) calls this essential character of cultural data the humanistic coefficient, “because such data, as objects

of the student's theoretical reflection, already belong to someone else's active experience and are such as this active experience makes them" (p. 136). Thus, if this cultural system was ignored and not seen as a humanistic coefficient, if cultural life was studied like a natural system, then the researcher would simply find "a disjointed mass of natural things and processes, without any similarity to the reality he started to investigate" (p. 137). It becomes central to research into human and cultural life.

Thus, one major tradition of social science has repeatedly stressed the importance of studying this "humanistic coefficient"; of getting at the ways in which participants of historical social life construct and make sense of their particular world—their "definitions of the situation," their "first level constructs." This tradition is exemplified in Thomas and Znaniecki's classic volume *The Polish Peasant in Europe and America*, where a whole array of subjective documents—letters, life stories, case records, and so on—are used to get at the meaning of the experience of migration. This major study anticipated the important development of personal documents, life histories, and participant observation within social science. These are among the core methods that help us get at the humanistic coefficient.

—Ken Plummer

See also VERSTEHEN, SENSITIZING CONCEPTS, LIFE HISTORY METHOD, PARTICIPANT OBSERVATION

REFERENCES

- Bryun, T. S. (1966) *The human perspective in sociology: The methodology of participant observation*. Englewood Cliffs, NJ: Prentice Hall.
- Thomas, W. I., & Znaniecki, F. (1918–1920). *The Polish peasant in Europe and America* (Vols. 1 & 2). New York: Dover.
- Znaniecki, F. (1934). *The method of sociology*. New York: Farrar and Rinehart.
- Znaniecki, F. (1969). *On humanistic sociology: Selected papers* (R. Bierstadt, Ed.). Chicago: University of Chicago Press.

HYPOTHESIS

This is a central feature of scientific method, specifically as the key element in the hypothetico-deductive model of science (Chalmers, 1999). Variations on this

model are used throughout the sciences, including social science. Hypotheses are statements derived from an existing body of THEORY that can be tested using the methods of the particular science. In chemistry or psychology, this might be an experiment, and in sociology or political science, the social survey. Hypotheses can be confirmed or falsified, and their status after testing will have an impact on the body of theory, which may be amended accordingly and new hypotheses generated. Consider a simplified example: Classical migration theory states that economic push and pull factors will be the motivation for migration, and agents will have an awareness of these factors. From this, it might be hypothesized that migrants will move from economically depressed areas to buoyant ones. If, upon testing, it is found that some economically depressed areas nevertheless have high levels of immigration, then the original theory must be amended in terms of either its claims about economic factors, or the agents' knowledge of these, or possibly both.

Rarely, however, is it the case that hypotheses are proven wholly right (confirmed) or wholly wrong (falsified), and even rarer that the parent theory is wholly confirmed or falsified. Through the second half of the 20th century, there was a great deal of debate (which continues) around the degree to which hypotheses and, by implication, theories can be confirmed or falsified. Karl Popper (1959) maintained that, logically, only a falsification approach could settle the matter, for no matter how many confirming instances were recorded, one can never be certain that there will be no future disconfirming instances. Just one disconfirming instance will, however, demonstrate that something is wrong in the specification of the theory and its derived hypothesis. Therefore, in this view, scientists should set out to *disprove* a hypothesis.

Falsification is logically correct, and its implied skepticism may introduce rigor into hypothesis testing, but in social science in particular, hypotheses cannot be universal statements and are often probabilistic. In the example above, no one would suggest that everyone migrates under the particular given circumstances, but that one is more or less likely to migrate. Indeed, in many situations, the decision as to whether one should declare a hypothesis confirmed or falsified may be just a matter of a few percentage points' difference between attitudes or behaviors in survey findings.

Hypotheses can be specified at different levels. Research hypotheses are linguistic statements about

the world whose confirmation/falsification likewise can be stated only linguistically as “none,” “all,” “some,” “most,” and so on. In qualitative research, a hypothesis might be framed in terms of a social setting having certain features, which, through observation, can be confirmed or falsified. However, in survey or experimental research, hypothesis testing establishes the statistical significance of a finding and, thus, whether that finding arose by chance or is evidence of a real effect. A null hypothesis is stated—for example, that there is no relationship between migration and housing tenure—but if a significant relationship is found, then the null hypothesis is rejected and the alternative hypothesis is confirmed. The alternative hypothesis may be the same as, or a subdivision of, a broader research hypothesis (Newton & Rudestam, 1999, pp. 63–65).

—Malcolm Williams

REFERENCES

- Chalmers, A. (1999). *What is this thing called science?* (3rd ed.). Buckingham, UK: Open University Press.
- Newton, R., & Rudestam, K. (1999). *Your statistical consultant: Answers to your data analysis questions*. Thousand Oaks, CA: Sage.
- Popper, K. (1959). *The logic of scientific discovery*. London: Routledge.

HYPOTHESIS TESTING. See SIGNIFICANCE TESTING

HYPOTHETICO-DEDUCTIVE METHOD

This method is commonly known as the “method of hypothesis.” It involves obtaining or developing a THEORY, from which a hypothesis is logically deduced (deduction), to provide a possible answer to a “why” RESEARCH QUESTION associated with a particular research problem. The hypothesis is tested by comparing it with appropriate DATA from the context in which the problem is being investigated.

The foundations of the hypothetico-deductive method were laid by the English mathematician and theologian William Whewell (1847) as a counter to

the method of INDUCTION advocated much earlier by Francis Bacon. Whewell debated the notion of induction with his contemporary, John Stuart Mill (1879/1947). He was critical of Mill’s view that scientific knowledge consists of forming generalizations from a number of particular observations, and he challenged the view that observations can be made without preconceptions. He rejected the idea that generalizing from observations is the universally appropriate scientific method and argued that hypotheses must be invented at an early stage in scientific research in order to account for what is observed. For him, observations do not make much sense until the researcher has organized them around some conception. In short, he shifted the source of explanations from observations to constructions in the mind of the scientist that will account for observed phenomena.

These conceptions may involve the use of concepts or phrases that have not been applied to these “facts” previously. In Kepler’s case, it was *elliptical orbit*, and for Newton, it was *gravitate*. Whewell argued that the process requires “inventive talent”; it is a matter of guessing several conceptions and then selecting the right one. He thought that it was impossible to doubt the truth of a hypothesis if it fits the facts well. In spite of this self-validation, he was prepared to put hypotheses to the test by making predictions and appropriate observations. It is on this latter point that Whewell anticipated Karl Popper’s FALSIFICATIONISM.

Popper (1972) was not particularly interested in the notion of hypotheses as organizing ideas. Rather, he was concerned with the view of science in which hypotheses, or conjectures, were produced as tentative answers to a research problem, and were then tested. According to Popper, in spite of the belief that science proceeds from observation to theory, to imagine that we can start with pure observation, as in POSITIVISM, without anything in the nature of a theory is absurd. Observations are always selective and occur within a frame of reference or a “horizon of expectations.” Rather than wait for regularities to impose themselves on us from our observations, according to Popper, we must actively impose regularities upon the world. We must jump to conclusions, although these may be discarded later if observations show that they are wrong.

Hence, Popper concluded that it is up to the scientist to invent regularities in the form of theories. However, the attitude must be critical rather than dogmatic. The dogmatic attitude leads scientists to

look for confirmation for their theories, whereas the critical attitude involves a willingness to change ideas through the testing and possible refutation of theories. He argued that we can never establish whether theories are, in fact, true; all we can hope to do is to eliminate those that are false. Good theories will survive critical tests.

—Norman Blaikie

REFERENCES

- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Popper, K. R. (1972). *Conjectures and refutation*. London: Routledge & Kegan Paul.

- Mill, J. S. (1947). *A system of logic*. London: Longman, Green. (Originally published in 1879.)
- Whewell, W. (1847). *The philosophy of the inductive sciences* (Vols. 1 & 2). London: Parker.

I

IDEAL TYPE

Ideal types are CONSTRUCTS used for data analysis in qualitative social research. The idea that ideal types can be used for systematic comparative analysis of historical data (including life-history data) was originally developed by Max Weber. Weber's methodology of ideal types has been adapted in ideal-type analysis to fulfill the requirements of multi-case study research. Through the use of ideal types, life history data are analyzed systematically on a case-by-case basis, arriving at structural explanations. Ideal-type analysis is a branch of interpretive social research dealing with narrative (textual) data materials that are organized in multiple case trajectories.

HISTORICAL DEVELOPMENT

The idea of ideal type was originally devised by Max Weber to explain sociologically individual cultural phenomena. That is, Weber (1904/1949) constructed ideal types as explanatory schemes when he wished to understand individual case material taken from the "infinite causal web" of social reality (p. 84). Alfred Schütz, another theorist who also used ideal types in his research, realized that not only are ideal types valid constructions for microsociological theory, but the social world itself is organized in ideal-type structures (Schütz, 1932/1967, 1955/1976).

Against this background, Gerhardt (1985, 1994) introduced ideal-type analysis into qualitative research. She applied it in two major studies, demonstrating its systematic scope (Gerhardt, 1986, 1999). The

approach has become established in Germany and elsewhere (e.g., Leisering & Leibfried, 1999).

LEVELS OF CASE ANALYSIS

Cases are the focus of ideal-type analysis in three stages of the research act. For one, cases are units of analysis established through data processing; second, cases are selected for their relative capacity of ideal-type representation; and third, case explanation is the ultimate goal of ideal-type analysis. The three levels of case work are the following:

1. *Case reconstruction*, which entails reconstruction of all cases in the data set as a sequence of stages following an analytical scheme. Case reconstruction arranges cases as sequential patterns such that each case can be compared with all other cases in the data set.

2. *Selection of paradigmatic cases*, which is the outcome of comparative analysis of cases (following systematic case reconstruction). When clusters of similar cases emerge, within each cluster, one or more paradigmatic cases can be chosen that epitomize the respective typical pattern.

3. *Case explanation*, which is the eventual objective of the research act, follows structural explanation (see next section). For Weber and also for modern ideal-type analysis, the aim is to explain systematically the developmental dynamics of empirical cases.

ANALYTICAL PROCEDURE

Through analytical procedures on two separate levels, on both of which ideal types are used,

descriptions as well as structural explanations are reached. The descriptions serve to construct descriptive types that epitomize the dynamics of the cases. The explanations are geared toward understanding the structural patterns that explain the dynamics of the case material. The use of ideal types to understand patterns of social life on both a descriptive as well as a structural level has two separate stages in the analytical procedure:

1. *Descriptive Ideal Types.* Through comparison between cases as they are reconstructed using the analytical scheme, clusters of similar cases emerge. These may be pictured choosing a paradigmatic case taken as ideal type. The ideal-type case epitomizes the case dynamics in the respective cluster. Through comparison between a paradigmatic case (ideal-type case) and all other cases in the cluster, applying the same procedure to each cluster of cases, a rich picture of case dynamics emerges. This picture shows the pattern in each group as well as that of all groups on a comparative basis. This allows for a variety of comparisons in the empirical material, yielding a tableau of variation of patterns epitomized in ideal-type paradigmatic cases.

2. *Structural Ideal Types.* If the question is asked, “Why do cases develop as they do?” social structures come into the picture. Structural explanation of pattern dynamics is needed. Only through ideal-type-based structural explanation can explanation of individual cases be accomplished. Such explanation determines why a case follows a structural pattern dynamics, which it does more or less closely. To arrive at the selection of ideal-type cases that can be used for structural pattern explanation, a catalogue of case characteristics must be set up that—under the analytical perspective of the research project—defines what would be the features or composition of an optimal case. The ideal-type case chosen on this basis is to be found in the empirical material. The ideal-type case epitomizes the structural pattern, the latter being mirrored in its case characteristics as closely as possible. In-depth analysis of the ideal-type case reveals the structural dynamics of the “pure” type in the structural pattern. Heuristically, to use a Weberian term, the ideal-type case represents the structural pattern for analytical purposes. Analytically, however, not only an optimal case but also a worst-case scenario should be established. Juxtaposing the analytical pictures derived from the best-case as well as the worst-case analytical scenarios yields a prolific tableau of patterns. These patterns define the structural backgrounds of potential case dynamics in the analyzed

case material. High- versus low-class status groups and multimarriage versus never-married family background, to name but two, are social structural patterns whose dynamics may be investigated using ideal-type analysis.

Ideal-type cases allow for explanations linking cases with structural patterns. Because social structures are realized to a variable degree in the cases in the data material, their typicality in relation to the ideal-type case(s) helps understand their structural versus individual developmental dynamics. Comparative analysis juxtaposing ideal-type cases with other cases in the data material allows also for analysis of subgroups, and so on, which further enriches the findings achieved through structural pattern explanation.

AN EXAMPLE

Our study of biographies of coronary artery bypass (CABS) patients investigated 60 cases (using 240 interviews). They were selected preoperatively on the basis that each fulfilled four known criteria (social and medical) predicting postoperative return to work. Slightly more than half the cases did eventually return to work postoperatively; the others preferred or were eventually pushed into early retirement. Case reconstruction yielded four patterns, namely successful revascularization/return to work, failure of revascularization/return to work, successful revascularization/early retirement, failure of revascularization/early retirement. These four groups were investigated further using descriptive ideal types (cases paradigmatic for each of the four patterns). In order to explain the case trajectories, all cases were recoded using four criteria of optimum outcome (postoperative absence of symptoms, improvement of income, absence of marital conflict over postoperative decision regarding return to work or retirement, satisfactory doctor-patient relationship). Only two cases fulfilled all four criteria—one a return-to-work case, and the other an early-retirement one. They turned out to differ with respect to a number of explanatory parameters, namely social class, age at the time of CABS, marital status, subjective perception of health status, and—expressed in their narratives—central life interest. Whereas the upper-class, older return-to-work case had a central life interest favoring his job and occupation, the lower-class, younger early-retirement case had a non-work-related central life interest (favoring his leisure and the prospect of old age without work stress). The two ideal-type cases were then used to explain the course of events in

all other cases, using cross-case comparative analysis (comparing biographical trajectories against the background of parameters such as, among others, social class). One main result of the study was the following: Return to work or early retirement after CABS depends more on type of central life interest than outcome of revascularization.

IDEAL TYPES AND MEASUREMENT

Max Weber (1904/1949) made it clear that ideal-type analysis aims at measurement. He expressed it as follows: “Such concepts are constructs in terms of which we formulate relationships by the application of the category of objective possibility. By means of this category, the adequacy of our imagination, oriented and disciplined by reality, is *judged*” (p. 93). Following on from Weber, and showing how Weber’s ideas of ideal-type methodology can be transformed into a method of modern qualitative research, Gerhardt (1994) writes about case explanation as measurement of the case development in a particular individual biography whose course was pitted against an ideal-type case in the respective data material:

This case explanation may serve as an example of how the individual case often represents chance and unforeseen hazards. These emerge as specific when the case is held against the ideal-type case. Apparently irrational elements, to be sure, explain the case by making understandable its unique dynamics—just as Weber suggested. (p. 117)

—Uta Gerhardt

REFERENCES

- Gerhardt, U. (1985). Erzählzeiten und Hypothesenkonstruktion. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 37, 230–256.
- Gerhardt, U. (1986). *Patientenkarrieren: Eine idealtypen-analytische Studie*. Frankfurt: Suhrkamp.
- Gerhardt, U. (1994). The use of Weberian ideal-type methodology in qualitative data interpretation: An outline for ideal-type analysis. *BMS Bulletin de Methodologie Sociologique* (International Sociological Association, Research Committee 33), 45, 76–126.
- Gerhardt, U. (1999). *Herz und Handlungsrationalität: Eine idealtypen-analytische Studie*. Frankfurt: Suhrkamp.
- Leisering, L., & Leibfried, S. (1999). *Time and poverty in Western welfare states*. New York: Cambridge University Press.
- Schütz, A. (1967). *Der sinnhafte Aufbau der sozialen Welt* [Phenomenology of the social world]. Evanston, IL: Northwestern University Press. (Originally published in 1932)
- Schütz, A. (1976). Equality and the meaning structure of the social world. In *Collected Papers II* (pp. 226–273). The Hague: Martinus Nijhoff. (Originally published in 1955)
- Weber, M. (1949). Die “Objektivität” sozialwissenschaftlicher und sozialpolitischer Erkenntnis [“Objectivity” in social science and social policy]. In *The methodology of the social sciences: Max Weber* (pp. 50–112). New York: Free Press. (Originally published in 1904)

IDEALISM

Idealism is a philosophical position that underpins a great many epistemological stances in social science. All of these share a common view that reality is mind dependent or mind coordinated. Although idealism might be seen as contrary to REALISM, which postulates a mind-independent reality, idealist thought spans a range of positions. These positions do not reject outright the existence of an external reality; rather, they point to the limits of our knowledge of it.

Indeed, one of the most important idealist philosophers, Kant, held that the only way we can conceive of ourselves as mind-endowed beings is in the context of existing in a world of space and time (Körner, 1955, Chap. 4). What is important is that we can know phenomena only through mind-dependent perception of it, we cannot ever know the thing in itself, what he called *noumena*. Kant’s *transcendental idealism* has been enormously influential both directly and indirectly, perhaps most importantly in the work of Max Weber. He shifted the issue of knowing away from the thing itself to a refinement of our instruments of knowing, specifically “ideal types.” Ideal types are not averages or the most desirable form of social phenomena, but ways in which an individual can reason from a shared rational faculty to a model or pure conceptual type that would exist if all agents acted perfectly rationally. Therefore, Weber’s methodology depends crucially on rationality as the product of minds (Albrow, 1990).

Hegel’s idealism is very different from Kant’s. Kantian philosophy emphasizes the epistemological through a refinement of the instruments of knowing, whereas Hegel emphasizes the ontological through a development of our understanding of mind. Unlike Kant’s noumena, Hegel believed mind (and minds as a universal category) could be known through a

dialectical process of contradiction and resolution—ironically, a method most famously associated with Marx, a materialist and realist!

Most expressions of idealist social science are not obviously or directly traceable to Kant or Hegel, but simply begin from the premise that the social world consists of ideas, that it exists only because we think it exists, and we reproduce it on that basis. The means of that reproduction is considered to be language, and thus, idealism underpins the “linguistic turn” in social science. An important originator, at least in the Anglophone world, was Peter Winch (1958/1990), who argued that rule-following in social life consisted of linguistic practice and, because language is subject to variation (and shapes culture accordingly), then to know a culture, one must come to know the language of that culture. Winch’s philosophy is an important constituent of ETHNOMETHODOLOGY and, less directly, more recent POSTMODERNIST approaches.

—Malcolm Williams

REFERENCES

- Albrow, M. (1990). *Max Weber’s construction of social theory*. London: Macmillan.
- Körner, S. (1955). *Kant*. Harmondsworth, UK: Penguin.
- Winch, P. (1990). *The idea of a social science and its relation to philosophy*. London: Routledge. (Originally published in 1958)

IDENTIFICATION PROBLEM

In the social sciences, empirical research involves the consistent ESTIMATION of POPULATION parameters from a prespecified system consisting of data, a parameterized model, and statistical MOMENT conditions. In this context, an identification problem is any situation where the PARAMETERS of the model cannot be consistently estimated. Systems that yield consistent estimates of the model’s parameters are *identified*, and those that do not are *not identified*. Sometimes, a system may produce consistent estimates of a subset of its parameters, in which case the identification problem is at the parameter level. Identification problems can be loosely differentiated into two classes: those caused by *insufficient moment conditions*, and those caused by an *overparameterization* of the model. However, there is a substantive degree of overlap in these classifications (an overparameterized model can be thought of

as having too few moment conditions, and vice versa). Simple theoretical and empirical examples of these phenomena are discussed below.

ORDINARY LEAST SQUARES (OLS) regression can be thought of as a solution to an identification problem caused by *insufficient moment conditions*. Consider the simple *deterministic* system

$$y_i = \alpha + x_i\beta + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where the ε_i are nonrandom slack terms. Equation (1) alone cannot determine the unknown parameters α and β ; there are essentially n equations and $n + 2$ unknowns (think of the n slack terms as unknowns, too). Therefore, the system lacks two equations for identification. However, if the slack terms are stochastic errors, then the two OLS moment conditions

$$E(\varepsilon_i) = 0 \text{ (zero-mean condition)}, \quad (2)$$

$$E(x_i\varepsilon_i) = 0 \text{ (exogeneity condition)} \quad (3)$$

provide the two necessary equations to identify consistent estimates of α and β via OLS regression.

The familiar INSTRUMENTAL VARIABLE technique is a leading case of a solution to an identification problem. If the exogeneity condition in equation (3) fails, then the system in equations (1) and (2) lacks one identifying moment condition. If there are suitable instruments z_i for the x_i , then the four additional equations

$$x_i = \gamma + z_i\delta + v_i, \quad i = 1, \dots, n,$$

$$E(v_i) = 0,$$

$$E(z_iv_i) = 0,$$

$$E(z_i\varepsilon_i) = 0,$$

along with equations (1) and (2), identify the parameters: α , β , γ , and δ (effectively, $2n + 4$ equations and $2n + 4$ unknowns, counting the ε_i and v_i as unknowns). Notice that the system defined by equations (1) and (2) is also *identified* (trivially) if $\beta = 0$, so this identification problem could be thought of naively as one of *overparameterization*.

A substantive example of the *overparameterization* class of identification problems is the Simultaneous Equations problem, which is typically solved with parameter restrictions on the system. Simultaneously determined variables in a system containing more than one modeling equation often fail exogeneity conditions like equation (3). For example, if y_i and x_i are

simultaneously determined and z_i is EXOGENOUS, then a simultaneous model might be

$$y_i = x_i\beta_1 + z_i\theta_1 + \varepsilon_{1i}, \quad (4)$$

$$x_i = y_i\beta_2 + z_i\theta_2 + \varepsilon_{2i}, \quad (5)$$

$i = 1, \dots, n$, where $E(\varepsilon_{1i}) = E(\varepsilon_{2i}) = E(\varepsilon_{1i}\varepsilon_{2i}) = E(z_i\varepsilon_{1i}) = E(z_i\varepsilon_{2i}) = 0$. Neither equation in this system is *identified*. To see this, notice that y_i in equation (4) is correlated with ε_{2i} through x_i in equation (5). Similarly, x_i in equation (5) is correlated with ε_{1i} through y_i in equation (4). Therefore, the entire system fails the OLS exogeneity condition and cannot be consistently estimated. Moreover, some algebra shows that equations (4) and (5) are identical up to parameters, so no solution is forthcoming. If theory suggests the parameter restriction $\theta_1 = 0$, then the equations are not identical, and z_i may be an appropriate instrument for y_i . Then, equation (5) could be consistently estimated using instrumental variable techniques. Equation (4) remains *not identified*, because all of the information contained in the data (x_i, y_i, z_i) has been exhausted. A complete theoretical treatment of identification problems in Simultaneous Equations models can be found in Greene (2003), Kmenta (1997), Schmidt (1976), or Wooldridge (2003).

Manski (1995) presents some empirical examples of the identification problem in applied social science research. For example, Manski's *mixing problem* in social program evaluation is an identification problem. That is, outcomes of controlled social program experiments, administered homogeneously to an experimental treatment sample, may not produce the same results that would exist were the social program to be administered heterogeneously across the population. Therefore, the potential outcomes of a social program across the population may not be identified in the outcomes of a social program experiment. Additionally, the *selection problem* in economic models of wage determination is an identification problem. Market wage data typically exclude wage observations of individuals in the population who choose not to work; therefore, estimates of wage parameters for the general population (working and nonworking), based solely on wage data from those that select into the labor market, may not be identified. In both of these examples, identification is achieved by bringing prior information to bear on the problem; see Manski (1995). Insofar as this prior information can be formulated

as additional moment conditions, the previous mixing and selection problems are identification problems of insufficient moment conditions.

—William C. Horrace

REFERENCES

- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Kmenta, J. (1997). *Elements of econometrics* (2nd ed.). Ann Arbor: University of Michigan Press.
- Manski, C. F. (1995). *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.
- Schmidt, P. (1976). *Econometrics*. New York: Marcel Dekker.
- Wooldridge, J. M. (2003). *Introductory econometrics* (2nd ed.). Mason, OH: South-Western.

IDIOGRAPHIC/NOMOTHETIC. See
NOMOTHETIC/IDIOGRAPHIC

IMPACT ASSESSMENT

Impact assessment involves a comprehensive evaluation of the long-term effects of an intervention. With regard to some phenomenon of interest, such as a community development program, the evaluator asks: What impact has this program had on the community, in terms of both intended and unintended effects (Patton, 1997)?

Impact assessment brings a broad and holistic perspective to evaluation. The first level of assessment in evaluation is at the resource or input level: To what extent did the project or program achieve the targeted and needed level of resources for implementation? The second level of assessment is implementation analysis: To what extent was the project or program implemented as planned? The third level of analysis focuses on outputs: To what extent did the project or program produce what was targeted in the original plan or proposal in terms of levels of participation, completion, or products produced? The fourth level of analysis involves evaluating outcomes: To what extent and in what ways were the lives of participants in a project or program improved in accordance with specified goals and objectives? Finally, the ultimate level of evaluation is impact assessment: To what extent, if at all, were outcomes sustained or increased over the long term, and what

ripple effects, if any, occurred in the larger context (e.g., community)?

Consider an educational program. The first level of evaluation, inputs assessment, involves looking at the extent to which resources are sufficient—adequate buildings, materials, qualified staff, transportation, and so forth. The second level of evaluation, implementation analysis, involves examining the extent to which the expected curriculum is actually taught and whether it is taught in a high-quality way. The third level of evaluation involves outputs, for example, graduation rates; reduced dropout rates; parent involvement indicators; and satisfaction data (student, parents, and teachers). The fourth level, outcomes evaluation, looks at achievement scores, quality of student work produced, and affective and social measures of student growth. Finally, impact assessment looks at the contribution of the education program to students after graduation, over the long term, and the effects on the broader community where the program takes place, for example, community pride, crime rates, home ownership, or businesses attracted to a community because of high-quality schools or repelled because of poorly performing schools.

In essence, then, there are two dimensions to impact assessment, a *time dimension* and a *scope dimension*. The time dimension is long term: What effects endure or are enhanced beyond immediate and shorter term outcomes? *Outcomes evaluation* assesses the direct instrumental linkage between an intervention and participant changes like increased knowledge, competencies, and skills, or changed attitudes and behaviors. In contrast, *impact assessment* examines the extent to which those outcomes are maintained and sustained over the long term. For example, a school readiness program may aim to prepare young children for their first year of school. The outcome is children's performance during that first year in school. The impact is their performance in subsequent years.

The second dimension of impact assessment is scope. Outcomes evaluation looks at the narrow, specific, and direct linkage between the intervention and the outcome. Impact assessment looks more broadly for ripple effects, unintended consequences, side effects, and contextual effects. Impact assessment includes looking for both positive and negative effects, as well as both expected and unanticipated effects. For example, the desired and measured outcome of a literacy program is that nonliterate adults learn to read. The impacts might include better jobs for those

adults, greater civic participation, and even higher home ownership or lower divorce rates. Therefore, impact assessment looks beyond the direct, immediate causal connection to longer term and broader scope effects.

This broader and longer term perspective creates methodological challenges. Everything and anything is open for investigation because one is seeking unexpected and unanticipated effects as well as planned and foreseen ones. The community is often the unit of analysis for impact assessments (Lichfield, 1996). Comprehensive impact assessment requires multiple methods and measurement approaches, both QUANTITATIVE and QUALITATIVE, and both deductive and inductive reasoning. CAUSAL connections become increasingly diffuse and difficult to follow over time and as one moves into more complex systems beyond the direct and immediate causal linkages between a program INTERVENTION and its immediate outcome. Therefore, impact assessment is also costly.

With greater attention to the importance of systems understandings in social science, impact assessment has become increasingly valued, although challenging to actually conduct.

—Michael Quinn Patton

REFERENCES

- Lichfield, N. (1996). *Community impact evaluation*. London: University College Press.
 Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

IMPLICIT MEASURES

The distinction between *implicit* and *explicit* processes originated from work in cognitive psychology but is now popular in other areas, such as social psychology. Implicit processes involve lack of awareness and are unintentionally activated, whereas explicit processes are conscious, deliberative, and controllable. Measures of implicit processes (i.e., implicit measures) include response latency procedures (e.g., the Implicit Association Test) (Greenwald et al., 2002); memory tasks (e.g., the process-dissociation procedure) (Jacoby, Debnier, & Hay, 2001); physiological reactions (e.g., galvanic skin response); and indirect questions (e.g., involving attributional biases).

Implicit measures include such diverse techniques as the tendency to use more abstract versus concrete language when describing group members, preferences for letters that appear in one's own name, and response times to make decisions about pairs of stimuli. In one popular response-time technique for assessing racial biases, the Implicit Association Test (IAT) (Greenwald et al., 2002), participants first perform two categorization tasks (e.g., labeling names as Black or White, classifying words as positive or negative) separately, and then the tasks are combined. Based on the well-substantiated assumption that people respond faster to more associated pairs of stimuli, the speed of decisions to specific combinations of words (e.g., Black names with unpleasant attributes vs. Black names with unpleasant characteristics) represents the implicit measure. In contrast, explicit measures are exemplified by traditional, direct, SELF-REPORT MEASURES.

Perhaps because the various types of implicit measures are designed to address different aspects of responding (e.g., awareness or controllability) and focus on different underlying processes or systems (e.g., conceptual or perceptual), relationships among different implicit measures of the same social concepts or attitudes are only moderate in magnitude and highly variable. In addition, implicit measures have somewhat lower levels of TEST-RETEST RELIABILITY than do explicit measures (Fazio & Olson, 2003).

The relationship between implicit and explicit measures is also modest and variable (Dovidio, Kawakami, & Beach, 2001). One potential reason for the discordance between implicit and explicit measures is that they are differentially susceptible to strategic control. Explicit measures are more transparent and higher in REACTIVITY than are implicit measures. As a consequence, the correspondence between implicit and explicit measures tends to be weaker in more socially sensitive domains. Another explanation for the inconsistent relation between implicit and explicit measures is that people often have dual attitudes about people and objects—one relatively old and habitual, the other newer and conscious. Because implicit measures tap habitual attitudes and explicit measures assess more recently developed attitudes, implicit and explicit measures may be uncorrelated when the two components of the dual attitudes diverge.

In conclusion, implicit measures represent a diverse range of techniques that is fundamentally different from direct, explicit measures in terms

of objectives, implementation, assumptions, and interpretation. Moreover, because these are newly emerging techniques, their psychometric properties (RELIABILITY and VALIDITY) tend to be less well substantiated than are those for many established explicit measures (e.g., prejudice scales). Nevertheless, the application of implicit measures has already made substantial contributions for understanding human memory, cognition, and social behavior.

—John F. Dovidio

REFERENCES

- Dovidio, J. F., Kawakami, K., & Beach, K. R. (2001). Implicit and explicit attitudes: Examination of the relationship between measures of intergroup bias. In R. Brown & S. L. Gaertner (Eds.), *Blackwell handbook of social psychology: Intergroup processes* (pp. 175–197). Malden, MA: Blackwell.
- Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, 54, 297–327.
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109, 3–25.
- Jacoby, L. L., Debner, J. A., & Hay, J. F. (2001). Proactive interference, accessibility bias, and process dissociations: Valid subject reports of memory. *Journal of Experimental Psychology: Memory, Learning, and Cognition*, 27, 686–700.

IMPUTATION

MISSING DATA are common in practice and usually complicate data analyses. Thus, a general-purpose method for handling missing values in a data set is to impute, that is, fill in one or more plausible values for each missing datum so that one or more completed data sets are created. In principle, each of these data sets can be analyzed using standard methods that have been developed for complete data. Typically, it is easier first to impute for the missing values and then use a standard, complete-data method of analysis than to develop special statistical techniques that allow the direct analysis of incomplete data. In social surveys, for example, it is quite typical that the missing values occur on different variables. Income information might be most likely to be withheld, perhaps followed by questions about sexual habits, hygiene, or religion.

Correct analyses of the observed incomplete data can quickly become cumbersome. Therefore, it is often easier to use imputation.

Another great advantage of imputation occurs in the context of producing a public-use data set, because the imputer can use auxiliary confidential and detailed information that would be inappropriate to release to the public. Moreover, imputation by the data producer solves the missing data problem in the same way for all users, which ensures consistent analyses across users.

Imputation also allows the incorporation of observed data that are often omitted from analyses conducted by standard statistical software packages. Such packages often handle incomplete data by LISTWISE DELETION, that is, by deleting all cases with at least one missing item. This practice is statistically inefficient and often leads to substantially biased inferences. Especially in multivariate analyses, listwise deletion can reduce the available data considerably, so that they are no longer representative of the population of interest. A multivariate analysis based only on the completely observed cases will often lose a large percentage, say 30% or more, of the data, which, minimally, reduces the precision of estimates.

Moreover, basing inferences only on the complete cases implicitly makes the assumption that the missing data are missing completely at random (MCAR), which is not the case in typical settings. In social surveys, special socioeconomic groups or minorities are often disproportionately subject to missing values. If, in such cases, the missingness can be explained by the observed rather than missing variables, the missing data are said to be missing at random (MAR) and can be validly imputed conditionally on the observed variables. Often, it is quite plausible to assume that differences in response behavior are influenced by completely observed variables such as gender, age group, living conditions, social status, and so on.

If the missingness depends on the missing values themselves, then the data are missing not at random (MNAR). (See Little & Rubin, 2002, for formal definitions.) This might be the case for income reporting, where people with higher incomes tend to be less likely to respond, even within cells defined by observed variables (e.g., gender and district of residence). Even in such cases, however, (single or multiple) imputation can help to reduce bias due to nonresponse.

SINGLE IMPUTATION

Typically, imputation is used for item-NONRESPONSE, but it can also be used for unit-nonresponse. Many intuitively appealing approaches have been developed for single imputation, that is, imputing one value for each missing datum. A naïve approach replaces each missing value on a variable with the unconditional sample mean of that variable or the conditional sample mean after the cases are grouped (e.g., according to gender). In regression imputation, a REGRESSION of the variable with missing values on other observed variables is estimated from the complete cases. Then, the resulting prediction equation is used to impute the estimated conditional mean for each missing value. In stochastic regression imputation, a random residual is added to the regression prediction, where the random error has variance equal to the estimated residual variance from the regression.

Another common procedure is hot deck imputation, which replaces the missing values for an incomplete case by the observed values from a so-called donor case. A simple hot deck defines imputation cells based on a cross-classification of some variables that are observed for both complete and incomplete cases (e.g., gender and district of residence). Then, the cases with observed values are matched to each case with missing values to create a donor pool, and the observed values from a randomly chosen donor are transferred (imputed) to replace the missing values.

With nearest neighbor imputation, a distance (metric) on the observed variables is used to define the best donor cases. When the distance is based on the difference between cases on the predicted value of the variable to be imputed, the imputation procedure is termed *predictive mean matching*. In the past decade, considerable progress has been made in developing helpful hot deck and nearest neighbor procedures.

Little (1988) gives a detailed discussion of issues in creating imputations. One major consideration for imputation is that random draws rather than best predictions of the missing values should be used, taking into account all observed values. Replacing missing values by point estimates, such as means, conditional means, or regression predictions, tends to distort estimates of quantities that are not linear in the data, such as VARIANCES, COVARIANCES, and CORRELATIONS. In general, mean imputation can be very harmful to valid estimation and even more harmful to valid inference (such as CONFIDENCE INTERVALS). Stochastic

regression imputation and hot deck procedures are acceptable for point estimation, although special corrections are needed for the resulting variance estimates. Methods of variance estimation have been developed for the hot deck and nearest neighbor techniques, but they are limited to specific univariate statistics; for a recent discussion, see Groves, Dillman, Eltinge, and Little (2002).

MULTIPLE IMPUTATION

Imputing a single value for each missing datum and then analyzing the completed data using standard techniques designed for complete data will usually result in standard error estimates that are too small, confidence intervals that undercover, and p values that are too significant; this is true even if the modeling for imputation is carried out absolutely correctly. Multiple imputation (MI), introduced by Rubin in 1978 and discussed in detail in Rubin (1987), is an approach that retains the advantages of imputation while allowing the data analyst to make valid assessments of uncertainty. MI is a MONTE CARLO technique that replaces the missing values by $m > 1$ simulated versions, generated according to a probability distribution indicating how likely the true values are given the observed data. Typically, m is small, such as $m = 5$. Each of the imputed (and thus completed) data sets is first analyzed by standard methods; the results are then combined to produce estimates and confidence intervals that reflect the missing data uncertainty.

The theoretical motivation for multiple imputation is BAYESIAN, although the resulting multiple imputation inference is also usually valid from a frequentist viewpoint. Formally, Bayesian MI requires independent random draws from the posterior predictive distribution of the missing data given the observed data:

$$\begin{aligned} f(y_{\text{mis}}|y_{\text{obs}}) &= \int f(y_{\text{mis}}, \theta|y_{\text{obs}}) d\theta \\ &= \int f(y_{\text{mis}}|y_{\text{obs}}, \theta) f(\theta|y_{\text{obs}}) d\theta. \end{aligned}$$

(For ease of reading, the colloquial “distribution” is used in place of the formally correct “probability density function.”) Because it is often difficult to draw from $f(y_{\text{mis}}|y_{\text{obs}})$ directly, a two-step procedure for each of the m draws is useful:

1. Draw the parameter θ according to its observed-data posterior distribution $f(\theta|y_{\text{obs}})$.

2. Draw the missing data Y_{mis} according to their conditional predictive distribution, $f(y_{\text{mis}}|y_{\text{obs}}, \theta)$, given the observed data and the drawn parameter value from Step 1.

For many models, the conditional predictive distribution $f(y_{\text{mis}}|y_{\text{obs}}, \theta)$ in Step 2 is relatively straightforward due to iid aspects of the complete-data model used. On the contrary, the corresponding observed-data posterior distribution, $f(\theta|y_{\text{obs}}) = f(y_{\text{obs}}|\theta) f(\theta)/f(y_{\text{obs}})$, usually is difficult to derive, especially when the data are multivariate and have a complicated pattern of missingness. Thus, the observed-data posterior distribution is often not a standard distribution from which random draws can be generated easily. However, straightforward iterative imputation procedures have been developed to enable multiple imputation via MARKOV CHAIN MONTE CARLO METHODS; many examples are discussed extensively by Schafer (1997). Such data augmentation procedures (Tanner & Wong, 1987), which include, for example, GIBBS SAMPLING, yield a stochastic sequence $\{(y_{\text{mis}}^{(t)}, \theta^{(t)}), t = 1, 2, \dots\}$ whose stationary distribution is $f(y_{\text{mis}}, \theta|y_{\text{obs}})$. More specifically, an iterative sampling scheme is created, as follows. Given a current guess $\theta^{(t)}$ of the parameter, first draw a value $y_{\text{mis}}^{(t+1)}$ of the missing data from the conditional predictive distribution $f(y_{\text{mis}}|y_{\text{obs}}, \theta^{(t)})$. Then, conditioning on $y_{\text{mis}}^{(t+1)}$, draw a new value $\theta^{(t+1)}$ of θ from its complete-data posterior $f(\theta|y_{\text{obs}}, y_{\text{mis}}^{(t+1)})$. Assuming that t is suitably large, m independent draws from such chains can be used as multiple imputations of Y_{mis} from its posterior predictive distribution $f(y_{\text{mis}}|y_{\text{obs}})$.

Now consider the task of estimating an unknown quantity, say, Q . In the remainder of this section, the quantity Q , to be estimated from the multiply imputed data set, is distinguished from the parameter θ used in the model for imputation. Although Q could be an explicit function of θ , one of the strengths of the multiple imputation approach is that this need not be the case. In fact, Q could even be the parameter of a model chosen by the analyst of the multiply imputed data set, and this analyst’s model could be different from the imputer’s model. As long as the two models are not overly incompatible or the fraction of missing information is not high, inferences based on the multiply imputed data should still be approximately valid. Rubin (1987) and Schafer (1997, Chapter 4) and their references discuss the distinction between Q and θ more fully.

The MI principle assumes that the complete-data statistic, $\hat{Q}$, and its variance estimate, $\hat{V}(\hat{Q})$, can be regarded as the approximate complete-data posterior mean and variance for Q , that is, $\hat{Q} \approx E(Q|y_{\text{obs}}, y_{\text{mis}})$ and $\hat{V}(\hat{Q}) \approx V(Q|y_{\text{obs}}, y_{\text{mis}})$ based on a suitable complete-data model and prior distribution. Moreover, with complete data, tests and interval estimates based on the large sample normal approximation

$$(\hat{Q} - Q)/\sqrt{\hat{V}(\hat{Q})} \sim N(0, 1)$$

should work well. Notice that the usual MAXIMUM-LIKELIHOOD ESTIMATES and their asymptotic variances derived from the inverted Fisher information matrix typically satisfy these assumptions. Sometimes, it is necessary to transform the estimate $\hat{Q}$ to a scale for which the normal approximation can be applied; for other practical issues, see Rubin (1987), Schafer (1997) and Little and Rubin (2002).

After m imputed data sets have been created, the m complete-data statistics $\hat{Q}^{(i)}$ and their variance estimates $\hat{V}(\hat{Q}^{(i)})$ are easily calculated for $i = 1, \dots, m$. The final MI point estimate $\hat{Q}_{\text{MI}}$ for the quantity Q is simply the average

$$\hat{Q}_{\text{MI}} = \frac{1}{m} \sum_{i=1}^m \hat{Q}^{(i)}.$$

Its estimated total variance T is calculated according to the ANALYSIS OF VARIANCE principle, as follows:

- The “between-imputation” variance

$$B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q}^{(i)} - \hat{Q}_{\text{MI}})^2.$$

- The “within-imputation” variance

$$W = \frac{1}{m} \sum_{i=1}^m \hat{V}(\hat{Q}^{(i)}).$$

- The “total variance”

$$T = W + \left(1 + \frac{1}{m}\right) B.$$

The correction $(1 + \frac{1}{m})$ reflects the estimation of Q from a finite number (m) of imputations. Tests and two-sided interval estimates can be based on the approximation $(\hat{Q}_{\text{MI}} - Q)/\sqrt{T} \sim t_v$ where t_v denotes the Student’s t -distribution with degrees of freedom

$$v = (m-1) \left(1 + \frac{W}{(1+m^{-1})B}\right)^2.$$

Barnard and Rubin (1999) relax the above assumption of a normal reference distribution for the complete-data interval estimates and tests to allow a t -distribution, and they derive the corresponding degrees of freedom for the MI inference to replace the formula for v given here. Moreover, additional methods are available for combining vector estimates and covariance matrices, p values, and LIKELIHOOD RATIO STATISTICS (see Little & Rubin, 2002).

The multiple imputation interval estimate is expected to produce a larger but valid interval than an estimate based only on single imputation because the interval is widened to account for the missing data uncertainty and simulation error. Even with a small number of imputations (say, $m = 3$ or 4), the resulting inferences generally have close to their nominal coverages or significance levels. Hot deck procedures or stochastic regression imputation can be used to create multiple imputations, but procedures not derived within the Bayesian framework are often not formally proper in the sense defined by Rubin (1987), because additional uncertainty due to estimating the parameters of the imputation model is not reflected.

Proper MI methods reflect the sampling variability correctly; that is, the resulting multiple imputation inference is valid also from a frequentist viewpoint if the imputation model itself is valid. Stochastic regression imputation can be made proper by following the two-step procedure described previously for drawing Y_{mis} from its posterior predictive distribution, and thus, first drawing the parameters of the regression model (Step 1), and then using the drawn parameter values in the usual imputation procedure (Step 2). Rubin and Schenker (1986) propose the use of BOOTSTRAPPING to make hot deck imputation proper; the resulting procedure is called the approximate Bayesian bootstrap. This two-stage process first draws a bootstrap sample from each donor pool and then draws imputations randomly from the bootstrap sample; it is a nonparametric analogue to the two-step procedure for drawing Y_{mis} . Thus, approximately proper multiple imputations can often be created using standard imputation procedures in a straightforward manner.

COMPUTATIONAL ISSUES

Various single imputation techniques, such as mean imputation, conditional mean imputation, or regression imputation, are available in commercial statistical software packages such as SPSS. With the increasing computational power, more and more multiple imputation techniques are being implemented, making multiple

imputation procedures quite easy to implement for both creating multiply imputed data sets and analyzing them.

There are programs and routines available for free, such as the stand-alone Windows program NORM or the S-PLUS libraries NORM, CAT, MIX, PAN, and MICE (also available for R now). NORM uses a normal model for continuous data, and CAT a LOG-LINEAR MODEL for categorical data. MIX relies on a general location model for mixed categorical and continuous data. PAN is designed to address PANEL data. The new missing data library in S-PLUS 6 features these models and simplifies consolidating the results of multiple complete-data analyses after multiple imputation.

Moreover, there are the free SAS-callable application IVEware and the SAS procedures PROC MI and PROC MIANALYZE based on normal asymptotics. PROC MI provides a parametric and a nonparametric regression imputation approach, as well as the multivariate normal model. Both MICE and IVEware are very flexible tools for generating multivariate imputations for all kinds of variables by using chained equations. AMELIA (a free Windows version as well as a Gauss version) was developed especially to fit certain social sciences' needs.

Finally, SOLAS for Missing Data Analysis 3.0 is a commercial Windows program provided by Statistical Solutions Limited. For links and further details, see www.multiple-imputation.com or Horton and Lipsitz (2001).

SUMMARY AND SOME EXTENSIONS

Imputation enables the investigator to apply immediately the full range of techniques that is available to analyze complete data. General considerations and a brief discussion of modeling issues for creating imputations have been given; also discussed has been how multiply imputed data sets can be analyzed to draw valid inferences.

Another advantage of MI for the social sciences may lie in splitting long questionnaires into suitable parts, so that only smaller questionnaires are administered to respondents, promising better response rates, higher data quality, and reduced costs. The missing data are thus created by design to be MCAR or MAR, and can be multiply imputed with good results; for the basic idea, see Raghunathan and Grizzle (1995).

Although Bayesian methods are often criticized because prior distributions have to be chosen, this is also an advantage when IDENTIFICATION PROBLEMS

occur. For example, multiple imputations of missing data under different prior settings allow sensitivity analyses; for an application to statistical matching, see Rubin (1986) or Rässler (2002).

Finally, empirical evidence suggests that multiple imputation under MAR often is quite robust against violations of MAR. Even an erroneous assumption of MAR may have only minor impact on estimates and standard errors computed using multiple imputation strategies. Only when MNAR is a serious concern and the fraction of missing information is substantial does it seem necessary to model jointly the data and the missingness. Moreover, because the missing values cannot be observed, there is no direct evidence in the data to address the MAR assumption. Therefore, it can be more helpful to consider several alternative models and to explore the sensitivity of resulting inferences.

Thus, a multiple imputation procedure seems to be the best alternative at hand to account for missingness and to exploit all valuable available information. In general, it is crucial to use multiple rather than single imputation so that subsequent analyses will be statistically valid.

—Susanne Rässler, Donald B. Rubin,
and Nathaniel Schenker

AUTHORS' NOTE: The views expressed in this entry do not necessarily represent the views of the U.S. government, and Dr. Schenker's participation as a coauthor is not meant to serve as an official endorsement of any statement to the extent that such statement may conflict with any official position of the U.S. government.

REFERENCES

- Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948–955.
- Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (2002). *Survey nonresponse*. New York: Wiley.
- Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, 55, 244–254.
- Little, R. J. A. (1988). Missing data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287–301.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.
- Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54–63.
- Rässler, S. (2002). Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches. *Lecture Notes in Statistics*, 168. New York: Springer.

- Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4, 87–95.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366–374.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman and Hall.
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528–550.

INDEPENDENCE

Social scientists are usually interested in empirical questions such as, “Is living in a metropolitan area associated with facing a higher crime rate (Event B)?” and “Is making more than \$50,000 a year independent of having a college degree?” We answer these kinds of questions with reference to the concept of statistical *independence* and *dependence*. Two “events” are said to be statistically independent of each other if the occurrence of one event in no way affects the PROBABILITY of the other event. This general idea can be extended to any number of independent events that have nothing to do with each other. In contrast, if the occurrence of one event affects the likelihood of the other event, then these two events are said to be associated or statistically dependent.

If two events are independent from each other, the fact that one event has occurred will not affect the probability of the other event. Suppose neither the probability of Event A [i.e., $P(A)$] nor the probability of Event B [i.e., $P(B)$] is zero. Statistical independence between these two events suggests that the conditional probability of A given B [i.e., $P(A|B)$] is exactly the same as the marginal (or unconditional) probability of A [i.e., $P(A)$]. That is, if A and B are independent, then $P(A|B) = P(A)$. Likewise, one should expect that the conditional probability of B given A [i.e., $P(B|A)$] is the same as the probability of Event B [i.e., $P(B)$] if these two events are independent of each other. In other words, if A and B are

independent of each other, then $P(B|A) = P(B)$ and $P(A|B) = P(A)$.

The usual definition of statistical independence is that Event A and Event B are independent if and only if their joint probability [i.e., $P(A \cap B)$] is equal to the product of their respective marginal (or unconditional) probabilities, $P(A) \times P(B)$. Consider two independent Events A and B . We know $P(A|B) = P(A)$, then because $P(A|B) = \frac{P(A \cap B)}{P(B)}$, we obtain $P(A \cap B) = P(A) \times P(B)$ under the assumption that $P(B) \neq 0$. Likewise, the information $P(B|A) = P(B)$ leads to the same statement $P(B \cap A) = P(A) \times P(B)$ in the same way. This definition of mutual independence can be extended to multiple events. Specifically, Events $A, B, C, \dots, Z$, each having a nonzero probability, are independent if and only if

$$\begin{aligned} P(A \cap B \cap C \cap \dots \cap Z) \\ = P(A) \times P(B) \times P(C) \times \dots \times P(Z). \end{aligned}$$

Suppose we toss a fair six-sided die twice. Let Event A be the occurrence of an odd number and Event B be getting an even number. We know that $P(A) = \frac{3}{6} = \frac{1}{2}$ (1 or 3 or 5 occurs) and $P(B) = \frac{3}{6} = \frac{1}{2}$ (2 or 4 or 6 occurs). The probability that the first toss will show an odd number (Event A) and the second an even number (Event B) is $P(A \cap B) = P(A) \times P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ because the two tosses are independent from each other. Similarly, the probability of getting three tails from flipping an unbiased coin three times is simply $P(T_1 \cap T_2 \cap T_3) = P(T_1) \times P(T_2) \times P(T_3) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$.

Consider an example in which statistical independence does not hold. Table 1 is a CONTINGENCY TABLE showing population cross-classification of a purely hypothetical community. If a resident is randomly selected from this population, what is the probability that this resident will be a white Republican? Let A be the event that we get a Republican resident and B be the event that we get a white resident. Because there are 860 Republican residents in all, $P(A) = \frac{860}{1480}$ or about .581. Similarly, the probability of getting a white resident is $P(B) = \frac{1100}{1480}$ or about .743. Among the 1,100 white residents, there are 600 people who identified themselves as a Republican. Thus, if we are given the information that a resident is white, the probability of this person’s being a Republican is 600/1100 or about .545. Likewise, if we are given the information that a resident is a Republican, the probability of this person’s being a white resident is 600/860

Table 1 Population Cross-Classification Showing Statistical Dependence

Ethnic Group	Party Identification			Total
	Democratic	Independent	Republican	
White	400	100	600	1,100
Non-White	100	20	260	380
Total	500	120	860	1,480

or .698. We thus have $P(A) = .581$, $P(B) = .743$, $P(A|B) = .545$, and $P(B|A) = .698$. One can find that $P(A \cap B) \neq P(A) \times P(B)$ in this example because we obtain

$$P(A \cap B) = P(A) \times P(B|A) = .581 \times .698 \cong .406$$

$$\text{or} = P(B) \times P(A|B) = .743 \times .545 \cong .405.$$

Because $P(A) \times P(B) = .581 \times .743 = .432$, we find the joint probability of A and B , that is, $P(A \cap B)$, is not equal to the product of their respective probabilities. Thus, these two events are associated. Statistical methods employed to test independence between variables depend on the nature of the data. More information about the statistical test of independence may be found in Alan Agresti and Barbara Finlay (1997), Agresti (2002), and Richard A. Johnson and Gouri K. Bhattacharyya (2001).

—Cheng-Lung Wang

REFERENCES

- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley Interscience.
- Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences*. Upper Saddle River, NJ: Prentice Hall.
- Johnson, R. A., & Bhattacharyya, G. K. (2001). *Statistics: Principles and methods* (4th ed.). New York: Wiley.

INDEPENDENT VARIABLE

An independent variable is any variable that is held to have a causal impact on another variable, referred to as the DEPENDENT VARIABLE. The notion of what can be taken to be an independent variable varies between experimental and nonexperimental research (see entries on INDEPENDENT VARIABLE

(IN EXPERIMENTAL RESEARCH) and INDEPENDENT VARIABLE (IN NONEXPERIMENTAL RESEARCH).

—Alan Bryman

INDEPENDENT VARIABLE (IN EXPERIMENTAL RESEARCH)

A defining characteristic of every true experiment is the independent variable (IV), an adjustable or alterable feature of the experiment controlled by the researcher. Such variables are termed independent because they can be made independent of their natural sources of spatial and temporal covariation. The researcher decides whether or not the IV is administered to a particular participant, or group of participants, and the variable's magnitude or strength. Controlled variations in the IV are termed experimental treatments or manipulations. Variations in the presence or amount of an IV are termed the levels of the treatment. Treatment levels can be quantitative (e.g., 5, 10, 35 mg of a drug) or qualitative (participant assigned to Workgroup A, B, or C). IVs may be combined in FACTORIAL DESIGNS, and their interactive effects assessed. The IV is critical in the LABORATORY EXPERIMENT because it forms the basis for the inference of causation. Participants assigned randomly to different conditions of an experiment are assumed initially equivalent. Posttreatment differences on the DEPENDENT VARIABLE between participants receiving different levels of the IV are causally attributed to the IV.

Crano and Brewer (2002) distinguish three general forms of IVs: social, environmental, and instructional treatments. Social treatments depend on the actions of people in the experiment, usually actors employed by the experimenter, whose behavior is scripted or controlled. In Asch's (1951) classic conformity studies, naïve participants made a series of comparative judgments in concert with others. In fact, these others were not naïve, but were accomplices who erred consistently on prespecified trials. In one condition, one accomplice whose erroneous report was different from that of his peers shattered the (incorrect) unanimity of the accomplice majority. Differences between unanimous and nonunanimous groups were interpreted as having been caused by controlled variations in the uniformity of accomplices' responses.

Environmental treatments involve the systematic manipulation of some feature(s) of the physical setting. In an attitude change experiment, for example, all participants may be exposed to the same persuasive communication, but some may receive the message under highly distracting circumstances, with noise and commotion purposely created for the experiment. Differences in susceptibility to the message between participants exposed under normal versus distracting conditions are interpreted causally, if participants were randomly assigned to the distraction or normal communication contexts, and the experimenter controlled the presence or absence of the distraction.

Instructional manipulations depend on differences in instructions provided to participants. For example, Zanna and Cooper (1974) gave participants a pill and suggested to some that they would feel tense and nervous, whereas others were informed that the pill would help them relax. In fact, the pill contained an inert substance that had no pharmacological effects. Differences in participants' subsequent judgments were attributed to variations in the instructional manipulation, the IV.

For strict experimentalists, factors that differentiate participants (e.g., sex, religion, IQ, personality factors), and other variables not under the control of the researcher (e.g., homicide rates in Los Angeles), are not considered independent and thus are not interpreted causally. However, in some research traditions, variables not under experimental control sometimes are suggested as causes. For example, a strong negative correlation between marriage rates in a society at Time 1, and suicides at Time 2, might be interpreted as marriage causing a negative impact on suicides. Marriage, it might be suggested, causes contentment and thus affects the likelihood of suicide. The opposite conclusion is untenable: Suicide rates at Time 2 could not have affected earlier marriage rates. However, such conclusions are usually (and appropriately) presented tentatively, insofar as a third, unmeasured variable might be the true cause of the apparent relationship. In the present example, economic conditions might be the true causal operator, affecting marriage rates at Time 1, and suicides at Time 2. Owing to the possibility of such third-variable causes, causal inferences based on correlational results are best offered tentatively. In the ideal experimental context, such extraneous influences are controlled, and hence their effects cannot be used to explain differences between groups receiving different levels of the IVs.

—William D. Crano

REFERENCES

- Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership, and men* (pp. 177–190). Pittsburgh, PA: Carnegie Press.
- Crano, W. D., & Brewer, M. B. (2002). *Principles and methods of social research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Zanna, M. P., & Cooper, J. (1974). Dissonance and the pill: An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology*, 29, 703–709.

INDEPENDENT VARIABLE (IN NONEXPERIMENTAL RESEARCH)

Also known as EXPLANATORY VARIABLE or input variable, an independent variable—or, more precisely, its effect on a DEPENDENT VARIABLE—is what a researcher analyzes in typical quantitative social science research. Strictly speaking, an independent variable is a variable that the researcher can manipulate, as in EXPERIMENTS. However, in nonexperimental research, variables are not manipulated, so that it is sometimes unclear which among the various variables studied can be regarded as INDEPENDENT VARIABLES. Sometimes, researchers use the term loosely to refer to any variables they include on the right-hand side of a REGRESSION equation, where causal inference is often implied. But it is necessary to discuss what qualifies as an independent variable in different RESEARCH DESIGNS. When data derive from a CROSS-SECTIONAL DESIGN, information on all variables is collected coterminously, and no variables are manipulated. This means that the issue of which variables have causal impacts on other variables may not be immediately obvious. Nonmanipulation of variables in disciplines such as sociology, political science, economics, and geography arises for several reasons: It may not be possible to manipulate certain variables (such as ethnicity or gender); it may be impractical to manipulate certain variables (such as the region in which people live); or it may be ethically and politically unacceptable (such as the causal impact of poverty). Some of these variables (e.g., race, ethnicity, gender, country of origin, etc.) are regarded as “fixed” in the analysis. Contrary to the definition of manipulability in experimental research, fixed variables are not at all manipulable but nevertheless are the ones that can be safely considered and treated as “independent” in nonexperimental research.

With longitudinal designs, such as a PANEL DESIGN, the issue is somewhat more complex, because although variables are still not manipulated, the fact that data can be collected at different points in time allows some empirical leverage on the issue of the temporal and hence causal priority of variables. Strict experimentalists might still argue that the lack of manipulation casts an element of doubt over the issue of causal priority. To deal with this difficulty, analysts of panel data resort to the so-called cross-lagged panel analysis for assessing causal priority among a pair of variables measured at two or more points in time. The line between dependent and independent variables in such analysis in a sense is blurred because the variable considered causally precedent is treated as an independent variable in the first regression equation but as a dependent variable in the second. Conversely, the variable considered causally succedent is treated as dependent in the first equation and as independent in the second. Judging from analyses like this, it appears that the term *independent variable* can be quite arbitrary; what matters is clear thinking and careful analysis in nonexperimental research.

Given that it is well known that it is wrong to infer causality from findings from a single regression run about relationships between variables, which is what are usually gleaned from nonexperimental research, how can we have independent variables in such research? Researchers employing nonexperimental research designs (such as a SURVEY design) must engage in *causal inference* to tease out independent variables. Essentially, this process entails a mixture of commonsense inferences from our understanding about the nature of the social world and from existing theory and research, as well as analytical methods in the area that is the focus of attention. It is this process of causal inference that lies behind such approaches as CAUSAL MODELING and PATH ANALYSIS, to which the cross-lagged panel analysis is related.

To take a simple example: It is well known that age and voting behavior are related. Does that mean that all we can say is that the two variables are correlated? Because the way we vote cannot be regarded in any way as having a causal impact on our ages, it seems reasonable to assume that age is the independent variable. Of course, there may be INTERVENING VARIABLES or MODERATING VARIABLES at work that influence the causal path between age and voting behavior, but the basic point remains that it is still possible and sensible to make causal inferences about such variables.

Making causal inferences becomes more difficult when causal priority is less obvious. For example, if we find as a result of our investigations in a large firm that employees' levels of organizational commitment and their job satisfaction are correlated, it may be difficult to establish direction of causality, although existing findings and theory relating to these variables might provide us with helpful clues. A panel study is often recommended in such circumstances as a means of disentangling cause and effect.

—Alan Bryman and Tim Futing Liao

REFERENCES

- Davis, J. A. (1985). *The logic of causal order* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-055). Beverly Hills, CA: Sage
- Rosenberg, M. (1968). *The logic of survey analysis*. New York: Basic Books.

IN-DEPTH INTERVIEW

In-depth interview is a term that refers to the UNSTRUCTURED INTERVIEW, the SEMISTRUCTURED INTERVIEW, and, sometimes, to other types of INTERVIEW IN QUALITATIVE RESEARCH, such as NARRATIVE INTERVIEWING and LIFE STORY INTERVIEWING.

—Alan Bryman

INDEX

An index is a measure of an abstract theoretical CONSTRUCT in which two or more indicators of the construct are combined to form a single summary score. In this regard, a SCALE is a type of index, and, indeed, the two terms are sometimes used interchangeably. But whereas scales arrange individuals on the basis of patterns of attributes in the data, an index is simply an additive composite of several indicators, called items. Thus, a scale takes advantage of any intensity structure that may exist among the attributes in the data, whereas an index simply assumes that all the items reflect the underlying construct equally, and therefore, the construct can be represented by summing the person's score on the individual items.

Indexes are widely used by government agencies and in the social sciences, much more so than scales. The consumer price index, for example, measures prices in various areas of the economy, such as homes, cars, real estate, consumer goods, and so forth. All of these entities contribute to the average prices that consumers pay for goods and services and therefore are included in the index. This example indicates why an index is generally preferred over a single indicator. Thus, the price of gasoline might spike at a particular time, but the price of other entities may be stable. It would be inaccurate and misleading to base the consumer prices just on the increasing cost of gasoline when the other items remained unchanged.

Furthermore, an index may not be unidimensional. For example, political scientists use a 7-point index to measure party identification. This index is arranged along a continuum as follows: strong Democrat, weak Democrat, independent leaning toward Democrat, independent, independent leaning toward Republican, weak Republican, strong Republican. One can easily see that this index consists of two separate dimensions. The first is a directional dimension, from Democrat through independent to Republican. However, there also exists an intensity dimension, going from strong through weak, independent, weak, and back to strong.

Frequently, governments use indexes of official statistics. For example, the consumer price index is used as a measure of the level of prices that consumers pay, and the FBI's crime index is the sum of the seven so-called index crime rates and is used as an overall indicator of crime in the United States.

—Edward G. Carmines and James Woods

REFERENCES

- Carmines, E. G., & McIver, J. P. (1981). *Unidimensional scaling* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-024). Beverly Hills, CA: Sage.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

INDUCTION

In logic, induction is a process for moving from particular statements to general statements. This logic is used in the social sciences to produce THEORY from DATA. Many specific instances are used to produce a general conclusion that claims more than the evidence on which it is based. Theory consists of universal GENERALIZATIONS about connections between phenomena. For example, many observed instances of juvenile delinquents who have come from broken homes are used to conclude that "all juvenile delinquents come from broken homes."

This is a popular conception of how scientists go about their work, making meticulous and objective observations and measurements, and then carefully and accurately analyzing data. It is the logic of science associated with POSITIVISM, and it is contrasted with DEDUCTION, the logic associated with FALSIFICATIONISM.

The earliest form of inductive reasoning has been attributed to Aristotle and his disciples, known as *enumerative induction* or *naïve induction*. Later, in the 17th century, induction was advocated by Francis Bacon as *the scientific method*. He wanted a method that involved the "cleansing of the mind of all presuppositions and prejudices and reading the book of nature with fresh eyes" (O'Hear, 1989, p. 16). Bacon was critical of Aristotle's approach and argued that rather than simply accumulating knowledge by generalizing from observations, it is necessary to focus on negative instances. His was an "eliminative" method of induction that required the comparison of instances in which the phenomenon under consideration was both present and absent. Two centuries later, John Stuart Mill elaborated Bacon's ideas in his methods of agreement and difference. In the method of agreement, two instances of a phenomenon need to be different in all but one characteristic, whereas in the method of difference, two situations are required to be similar in all characteristics but one, the phenomenon being present with this characteristic.

Advocates of the logic of induction require that OBJECTIVE methods be used, that data be collected carefully and systematically, without preconceived ideas guiding their selection. Generalizations are then produced logically from the data. Once a generalization is established, further data can be collected to strengthen

INDICATOR. See INDEX

or verify its claim. Also, new observed instances can be explained in terms of the pattern in the generalization. For example, if the generalization about the relationship between juvenile delinquency and broken homes can be regarded as universal, then further cases of juvenile delinquency can be explained in terms of family background, or predictions can be made about juveniles from such a background. This is known as pattern explanation.

Induction has been severely criticized and rejected by many philosophers of science. The following claims have been contested: that preconceptions can be set aside to produce objective observations; that “relevant” observations can be made without some ideas to guide their selection; that inductive logic has the capacity to mechanically produce generalizations from data; that universal generalizations can be based on a finite number of observations; and that establishing patterns or regularities is all that is necessary to produce explanations (see Blaikie, 1993, Chalmers, 1982, and O’Hear, 1989, for critical reviews). DEDUCTION has been offered as an alternative.

Induction was first advocated as *the* method in the social sciences by Emile Durkheim in *Rules of Sociological Method* (1964) and was put into practice in his study of *Suicide*. In spite of its shortcomings, a modified form of induction is essential in descriptive social research concerned with answering “what” RESEARCH QUESTIONS (see Blaikie, 2000).

—Norman Blaikie

REFERENCES

- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Blaikie, N. (2000). *Designing social research: The logic of anticipation*. Cambridge, UK: Polity Press.
- Chalmers, A. F. (1982). *What is this thing called science?* St. Lucia: University of Queensland Press.
- Durkheim, E. (1964). *Rules of sociological method*. Glencoe, IL: Free Press.
- O’Hear, A. (1989). *An introduction to the philosophy of science*. Oxford, UK: Clarendon.

important examples of difference among individuals include income, wealth, power, life expectancy, and mathematical ability. There are inequalities among individuals in each of these dimensions, and different populations have different distributions of various characteristics across individuals. It is often the case that there are also inequalities among subgroups within a given population: men and women, rural and urban, black and white. It is often important to identify and measure various dimensions of inequality within a society. For the purposes of public policy, we are particularly interested in social inequalities that represent or influence differences in individuals’ well-being, rights, or life prospects.

How do we measure the degree of inequalities within a POPULATION? There are several important alternative approaches, depending on whether the DISTRIBUTION of the factor is approximately NORMAL across the population. If the factor is distributed approximately normally, then the measurement of inequality is accomplished through analysis of the MEAN and STANDARD DEVIATION of the factor over the population. On this approach, we can focus on the descriptive statistics of the population with respect to the feature (income, life expectancy, weight, test scores): the shape of the distribution of scores, the mean and MEDIAN scores for the population, and the standard deviation of scores around the mean. The standard deviation of the VARIABLE across the population provides an objective measure of the degree of dispersion of the feature across the population. We can use such statistical measures to draw conclusions about inequalities across groups within a single population (“female engineers have a mean salary only 70% of that of male engineers”); conclusions about the extent of inequality in separate populations (“the variance of adult body weight in the United States is 20% of the mean, whereas the variance in France is only 12% of the mean”); and so forth.

A different approach to measurement is required for an important class of inequalities—those having to do with the distribution of resources across a population. The distribution of wealth and income, as well as other important social resources, is typically not statistically normal; instead, it is common to find distributions that are heavily SKEWED to the low end (that is, large numbers of individuals with low allocations). Here, we are explicitly interested in estimating the degree of difference in holdings across the population from bottom to top. Suppose we have a population of size N ,

INEQUALITY MEASUREMENT

Individuals differ in countless ways. And groups are characterized by different levels of inequality in all the ways in which individuals differ. Several

and each individual i has wealth W_i , and consider the graph that results from rank ordering the population by wealth. We can now standardize the representation of the distribution by computing the percent of total wealth owned by each percent of population. And we can graph the percent of cumulative wealth against percentiles of population. The resulting graph is the Lorenz distribution of wealth for the population. Different societies will have Lorenz curves with different shapes; in general, greater wealth inequality creates a graph that extends further to the southeast. Several measures of inequalities result from the technique of organizing a population in rank order with respect to ownership of a resource. The GINI COEFFICIENT and the ratio of bottom quintile to top quintile of property ownership are common measures of inequality used in comparative economic development.

—Daniel Little

REFERENCES

- Rae, D. W., & Yates, D. (1981). *Equalities*. Cambridge, MA: Harvard University Press.
 Sen, A. (1973). *On economic inequality*. New York: Norton.

INEQUALITY PROCESS

The Inequality Process (IP) is a MODEL of the process generating inequality of income and wealth. The IP is derived from the surplus theory of social stratification of economic anthropology. In this theory, success in competition for surplus—wealth net of production cost—concentrates surplus, explaining why evidence of substantial inequality first occurs in the same layer in archeological excavations as the earliest evidence of hunter/gatherers' acquiring agriculture and its surpluses. Gerhard Lenski (1966) extended the surplus theory from the increase in inequality at the dawn of agriculture to the waning of inequality since the Industrial Revolution. Lenski HYPOTHESIZED that more educated people retain more of their surplus.

In words, the IP is as follows:

People are randomly paired. Each competes for the other's wealth with a 50% chance to win. Each has two traits, wealth and omega, ω , the proportion of wealth lost in a loss. Winners take what losers lose. Repeat process.

Although winners gain, winning streaks are short. Wealth flows to those with smaller ω . Given Lenski's hypothesis, $(1 - \omega_i)$ is the IP's OPERATIONALIZATION of a worker at education level i .

The solution of the equations of IP competition shows that the IP's equilibrium DISTRIBUTION can be approximated by a two-PARAMETER gamma PROBABILITY DENSITY FUNCTION (pdf):

$$f_{it}(x) \equiv \frac{\lambda_{it}^{\alpha_i}}{\Gamma \alpha_i} x^{(\alpha_i-1)} e^{-(\lambda_{it}x)} \quad (1)$$

where

$\alpha_i \equiv$ shape parameter of individuals with ω_i ,

$$\begin{aligned} \alpha_i &\approx \frac{1 - \omega_i}{\omega_i} \\ \omega_i &\approx \frac{1}{1 + \alpha_i} \end{aligned} \quad (2)$$

$\lambda_{it} \equiv$ shape parameter of individuals with ω_i ,

$$\lambda_{it} \approx (1 - \omega_i) \frac{(\frac{w_{1t}}{\omega_1} + \frac{w_{2t}}{\omega_2} + \dots + \frac{w_{it}}{\omega_i} + \dots + \frac{w_{It}}{\omega_I})}{\bar{x}_t}, \quad (3)$$

$x \equiv$ wealth > 0 , considered as a random variable,

$\bar{x}_t \equiv$ unconditional mean of wealth at time t ,

$w_{it} \equiv$ proportion of population with ω_i at time t .

The IP reproduces (a) the higher mean earnings of the more educated, (b) the lower GINI of their earnings, and (c) the shapes of the distribution of earnings by level of education. The mean of x in $f_{it}(x)$ is

$$\bar{x}_{it} = \frac{\alpha_i}{\lambda_{it}} \approx \frac{\bar{x}_t}{\omega_i (\frac{w_{1t}}{\omega_1} + \dots + \frac{w_{It}}{\omega_I})} \quad (4)$$

and increases with smaller ω_i , holding the distribution of education, the w_{it} s, constant. The McDonald and Jensen expression for the gamma pdf's Gini is

$$\text{Gini ratio of } f_{it}(x) = \frac{\Gamma(\alpha_i + \frac{1}{2})}{\sqrt{\pi} \Gamma(\alpha_i + 1)}. \quad (5)$$

Given equation (2), Figure 1 shows how the Gini of x in $f_{it}(x)$ rises with greater ω_i . Empirically, the Gini of the earnings of the less educated is higher.

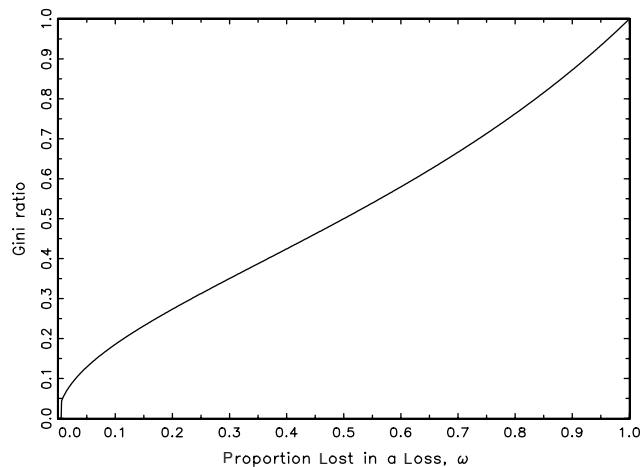


Figure 1 The Graph of the Gini Concentration Ratio of the Gamma PDF Whose Shape Parameter, α , Is Approximately Given by $(1 - \omega)/\omega$, Against ω

Figure 2, α_i , increases with education, so ω_i falls with education. The same sequence of shapes reappears throughout the industrial world.

A. B. Z. Salem and Timothy Mount (1974) advocate the two-parameter gamma pdf with no restrictions on its two parameters, the shape and scale parameters, to model income distribution. The parameters of the IP's gamma pdf model, $f_{it}(x)$, are restricted. Both are functions of ω_i . Figure 2 shows one year of the simultaneous fit of the IP's $f_{it}(x)$ to 33 years $\times$ 5 levels of education = 165 distributions. In the example, $f_{it}(x)$ has one parameter taking on five values in the population modeled, five parameter values to estimate. Salem and Mount's method requires estimating $(165 \times 2 =) 330$ parameters. Despite the fewer DEGREES OF FREEDOM, the IP's $f_{it}(x)$ fits nearly as well. Kleiber and Kotz (2003) discuss the IP in their chapter, "Gamma-Type Size Distributions".

—John Angle

Figure 2 shows $f_{it}(x)$ fitting the distribution of wage and salary income of nonmetropolitan workers in the United States in 1987 by level of education. The shape parameter of the gamma pdf fitting distributions in

REFERENCES

- Angle, J. (1986). The Surplus Theory of Social Stratification and the size distribution of personal wealth. *Social Forces*, 65, 293–326.

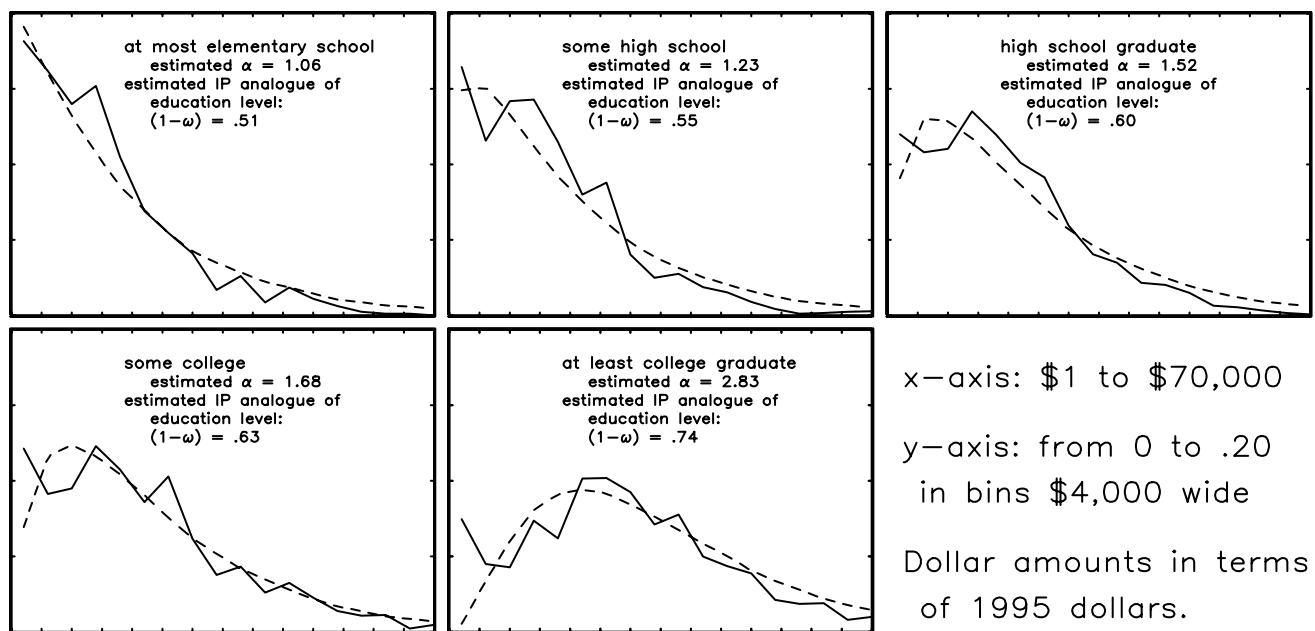


Figure 2 Relative Frequency Distribution of Annual Wage and Salary Income of Workers in 1987 (Residents of Nonmetropolitan Counties of U.S., Solid Curve) and Fitted IP Model (Dotted Curve). Workers Aged 25 to 65. One Year of Simultaneous Fits to Data from 1963 to 1995.

SOURCE: March Current Population Survey.

- Angle, J. (2002). The statistical signature of pervasive competition on wage and salary incomes. *Journal of Mathematical Sociology*, 26, 217–270.
- Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. Hoboken, NJ: Wiley.
- Lenski, G. (1966). *Power and privilege*. New York: McGraw-Hill.
- Salem, A., & Mount, T. (1974). A convenient descriptive model of income distribution: The gamma density. *Econometrica*, 42, 1115–1127.

INFERENCE. See STATISTICAL INFERENCE

INFERENCE. See STATISTICAL INFERENCE

INFERENCE. See STATISTICAL INFERENCE

Broadly, statistics may be divided into two categories: descriptive and inferential. Inferential statistics is a set of procedures for drawing conclusions about the relationship between sample statistics and population PARAMETERS. Generally, inferential statistics addresses one of two types of questions. First, what is the likelihood that a given sample is drawn from a population with some known parameter, θ ? Second, is it possible to estimate a population parameter, θ , from a sample statistic?

Both operations involve inductive rather than deductive logic; consequently, no closed form proof is available to show that one's inference is correct (see CLASSICAL INFERENCE). By drawing on the CENTRAL LIMIT THEOREM (CLT) and using the appropriate method for selecting our observations, however, making a probabilistic inference with an estimable degree of certainty is possible.

It can be shown empirically, and proven logically through the CLT, that if repeated random samples of a given size, n , are drawn from a population with a mean, μ , the sample means will be essentially normally distributed with the mean of the sample means approximating μ . Furthermore, if the population elements are distributed about the population mean with a VARIANCE of σ^2 , the sample means will have a sampling variance of σ^2/n . The standard deviation of the sample means is generally called the STANDARD ERROR to distinguish it from the standard deviation of a single sample.

As the sample size, n , increases, the closer will be the approximation of the distribution of the sample means to the normal distribution. Based on the description of the normal distribution, we know that

about 68% of the area (or data points) under the normal curve is found within ± 1 standard deviation from the mean. Because we are dealing with a distribution of sample means, we conclude that 68% of the sample means also will be within ± 1 standard error of the true population mean.

Since the normal distribution is ASYMPTOTIC, it is possible for a given sample mean to be infinitely far away from the population mean. Common sense, however, tells us that although this is possible, it is unlikely. Hence, we have the first application of inferential statistics. A random sample is selected, and a relevant sample statistic such as $\bar{y}$ is calculated. Often, a researcher wants to know whether $\bar{y}$ is equivalent to the known mean, μ , of a population. For example, a sociologist might want to know whether the average income of a random sample of immigrants is equal to that of the overall population. The sociologist starts with the null hypothesis that $\bar{y} - \mu = 0$ against the alternative that $\bar{y} - \mu \neq 0$. The value of $\bar{y}$ is calculated from the sample, and μ and σ^2 are obtained from census values. The quantity

$$z = \frac{\bar{y} - \mu}{\sqrt{\sigma^2/n}}$$

provides an estimate of how many standard deviations the sample mean is from the population parameter. As the difference $\bar{y} - \mu$ increases, the practical likelihood that $\bar{y} = \mu$ decreases, and the more likely it is that the sample is from a population different from the one under consideration. There is only a 5% chance that a sample mean two standard deviations away from the population mean (i.e., a z value larger than 2) would occur by chance alone; the likelihood that it would be three or more standard deviations away drops to less than 0.3%.

The second basic question addressed by inferential statistics uses many of the ideas outlined above but formulates the problem in a different way. For example, a polling company might survey a random sample of voters on election day and ask them how they voted. The pollster's problem is whether estimating the population value, μ , is possible based on the results of the survey. The general approach here is to generate a confidence interval around the sample estimate, $\bar{y}$. Using the sample variance, s^2 , as an estimate of the population variance, σ^2 , the pollster estimates the standard error for the population as $\sqrt{s^2/n}$.

The pollster decides that being 95% "certain" about the results is necessary. Thus, a 95% confidence

interval is drawn about $\bar{y}$ as: $\bar{y} \pm 1.96 \text{ s.e. } (\bar{y})$ or $\bar{y} \pm 1.96\sqrt{s^2/n}$. The pollster now concludes that μ would be within approximately two standard errors of $\bar{y}$ 19 times out of 20, or with a 95% level of confidence.

These two basic approaches can be elaborated to address a variety of problems. For example, we might want to compare two or more sample means to determine whether they are statistically equivalent. This is the same as asking whether all of the samples are members of the same population. We can modify the formulas slightly given the type of information that is available. For example, sometimes the population variance, σ^2 , is known, and sometimes it is not. When it is not known, we often estimate it from a sample variance. Similarly, when multiple samples are examined, we might pool their variances, if they are equivalent, to generate a better or more precise overall estimate of the population variance.

Clearly, one element that is central to inferential statistics is the assumption that the sampling distribution of the population is knowable. Using the CLT with nontrivial samples, we find that the sampling distribution approximates the normal distribution when simple random samples are selected. This assumption does not always hold. For example, with small samples, the population variance σ^2 is underestimated by the sample variance, s^2 . With small samples, we often use the t -distribution, which is flatter at the peak and thicker in the tails than the normal distribution. The t -distribution is actually a family of distributions. The shape of a specific distribution is determined by the degrees of freedom, which, in turn, is related to the sample size.

In other circumstances, yet other distributions are used. For example, with count data, we often use the chi-square distribution. In comparing variances, we use the F -distribution first defined by Fisher. Occasionally, the functional form of the sampling distribution may not be known. In those circumstances, we will sometimes use what is known as a nonparametric or distribution-free statistic. Social scientists use a variety of these procedures, including such statistics as GAMMA, the KOLMOGOROV-SMIRNOV test, and the various Kendall's tau statistics.

With the advent of inexpensive computing, more scientists are using modeling techniques such as the jackknife and BOOTSTRAPPING to estimate empirical distributions. These procedures are not always ideal but can be used successfully to estimate the sampling distributions for such statistics as the median or the INTERQUARTILE RANGE. Modeling techniques are

also useful when complex or nonrandom sampling procedures are used to generate samples.

There are also different frameworks within which the interpretation of results can be embedded. For example, BAYESIAN INFERENCE includes prior knowledge about the parameter based on previous research or the researcher's theoretical judgment as part of the inferential process. The likelihood approach, on the other hand, considers the parameter to have a distribution of possible values as opposed to being immutably fixed. Thus, within this approach, the distribution function of the sample results is judged conditional on the distribution of the possible parameter values.

—Paul S. Maxim

REFERENCES

- Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Hacking, I. (1965). *The logic of statistical inference*. Cambridge, UK: Cambridge University Press.
- Maxim, P. S. (1999). *Quantitative research methods in the social sciences*. New York: Oxford University Press.

INFLUENTIAL CASES

One or a few cases may be considered as influential observations in a sample when they have excessive effects on the outcomes of an analysis.

—Kwok-fai Ting

INFLUENTIAL STATISTICS

One or a few cases may be considered as influential observations in a sample when they have excessive effects on the outcomes of an analysis. In other words, the removal of these observations may change some aspects of the statistical findings. Thus, the presence of influential observations threatens the purpose of making generalizations from a wide range of data. For example, a few observations with extremely large or exceptionally small values make the mean a less useful statistic in describing the central tendency of a distribution. Besides parameter estimates, influential observations can also affect the validity of SIGNIFICANCE TESTS and GOODNESS-OF-FIT MEASURES in small and

large samples. Because of these potential problems, various influence statistics are formulated to assess the influence of each observation in an analysis.

Influence statistics were first discussed in linear regression analysis and were later extended to LOGISTIC REGRESSION analysis. These statistics attempt to measure two major components: leverage and residuals. Leverage is the power that an observation has in forcing the fitted function to move closer to its observed response. In regression analysis, observations that are further away from the center—that is, the point of the means of all predictors—have greater leverage in their favor. Thus, high-leverage observations are those with outlying attributes that separate themselves from the majority of a sample. A residual is the deviation of an observed response from the fitted value in an analysis. Large residuals are unusual in the sense that they deviate from the expected pattern that is projected by a statistical model. This is a cause of concern because observations with large residuals are often influential in the outcomes of a regression analysis; that is, the exclusion of these observations brings the fitted function closer to the remaining observations.

Graphic plots are useful visual aids for assessing the influence of outlying observations. Figure 1 illustrates three such observations that may have an undue influence on the fitted regression function in a bivariate analysis of hypothetical data. Observation 1 has a high leverage because it lies far away from the center of X , and it will exert a greater influence on the slope of the regression line. Although Observation 2 is near to the center of X , it deviates considerably from the expected value of the fitted regression function. The large residual is likely to affect the goodness-of-fit measure and significant test statistics. Observation 3 combines a high leverage with a large residual, making it more likely to distort the analytical results. For multivariate analyses, partial regression plots are more appropriate for the detection of influential observations because they take into account the multiple predictors in a model.

For quantitative measures, the hat-value is a common index of an observation's leverage in shaping the estimates of an analysis. Residuals are often measured by the studentized residuals—the ratios of residuals to their estimated standard errors. Several widely used influence statistics combine leverage and residuals in their computations. An idea that is common to these measures is the approximation of the consequences of sequentially dropping each observation to

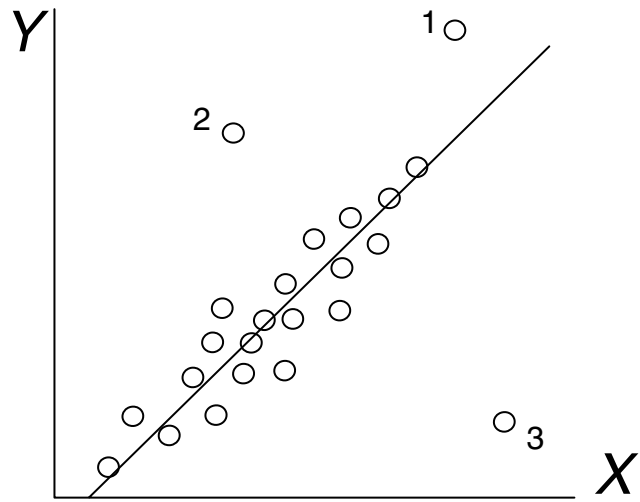


Figure 1 Bivariate Scatterplot With a Regression Line Fitted Without the Outlying Observations

assess its influence in an analysis. DFFITS examines the difference between two fitted response values of the same observation, with and without that observation contributing to the estimation. Cook's distance considers the effects of excluding an observation on all fitted values. DFBETAS evaluates the effects of excluding each observation on regression coefficients.

Influence statistics are heuristic devices that alert analysts of problematic observations in analyses. Although numeric cutoffs have been proposed for these statistics, many of them are suggestive rather than implied by statistical theories, and they should be used with flexibility. Graphic plots of these statistics are useful means of spotting major gaps that separate problematic observations from the rest. In practice, analysts need to replicate the analysis by excluding problematic observations to fully evaluate their actual effects. This does not mean that these observations, even when verified as influential to a given analysis, should be discarded in the final analysis. It is more important to seek the causes of their outlying patterns before looking for remedies. Common causes include coding errors, model misspecifications, and the pure chance of sampling.

—Kwok-fai Ting

REFERENCES

- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.

- Bollen, K. A., & Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. In J. Fox & J. S. Long (Eds.), *Modern methods of data analysis* (pp. 257–291). Newbury Park, CA: Sage.
- Collett, D. (1991). *Modelling binary data*. New York: Chapman and Hall.
- Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.
- Fox, J. (1991). *Regression diagnostics*. Newbury Park, CA: Sage.

INFORMANT INTERVIEWING

Informant interviews are of a more in-depth, less structured manner (semistructured, unstructured interviews) with a small selected set of informants most often in a field setting. What distinguishes this type of interview from other possible forms (e.g., cognitive tasks, survey interviews) is the depth, breadth, structure, and purpose of the interview format. Informant interviews can be used at various stages of the research enterprise to achieve a variety of different research interview objectives. Johnson and Weller (2002) make the distinction between top-down and bottom-up interviews. Bottom-up informant interviews aid in the clarification of less well-understood topics and knowledge domains and contribute to the construction of more valid STRUCTURED INTERVIEWING formats (e.g., structured questions) used in later stages of the research. Top-down informant interviews aid in the validation and alteration of existing structured questions (e.g., SURVEY questions) leading to the more valid adaptation of existing interviewing forms (e.g., cognitive tasks) to the research context.

An important use of informant interviews concerns the exploration of a less well-understood topic of interest. Informants are selected on the basis of their knowledge, experience, or understanding of a given topical area (Johnson, 1990). In such cases, the interviewer takes on the role of student and directs the interview in such a way as to learn from the informant, who is an expert in the area. Spradley's *The Ethnographic Interview* (1979) provides an excellent discussion of interviewing in this vein in his distinction of "grand tour" and "mini tour" questions in initial stages of interviewing. Grand tours involve asking questions that literally have the informant take the researcher on a tour of a given place, setting, or topical area. Other questions may direct informants,

for example, to describe a typical day or a typical interaction. Mini tour questions seek more specific information on any of these more general topics. The objective of these unstructured questions is to learn more about the informant's subjective understanding of social setting, scene, or knowledge domain.

An understanding of local terminology and issues gained in tour type questioning can facilitate the development of more specific questions for informants. There are a variety of more directed and structured informant interview formats that include, for example, taxonomic approaches or free recall elicitation that provide for a more organized and systematic understanding of a given topic on the part of the informants. Furthermore, such interviews can be used to develop more systematic types of questions and scales (i.e., bottom-up) that provide the ability for comparative analysis of beliefs, knowledge, and so on among informants (or respondents). In sum, informant interviews are an essential component of many forms of research, particularly ETHNOGRAPHIC research, in that they provide for the collection of data that contributes significantly to the overall VALIDITY of the research enterprise.

—Jeffrey C. Johnson

REFERENCES

- Johnson, J. C. (1990). *Selecting ethnographic informants*. Newbury Park, CA: Sage.
- Johnson, J. C., & Weller, S. C. (2002). Elicitation techniques for interviewing. In J. F. Gubrium & J. A. Holstein (Eds.), *The handbook of interview research*. Thousand Oaks, CA: Sage.
- Spradley, J. P. (1979). *The ethnographic interview*. New York: Holt, Rinehart and Winston.

INFORMED CONSENT

Voluntary informed consent refers to a communication process in which a potential subject decides whether to participate in the proposed research. *Voluntary* means without threat or undue inducement. *Informed* means that the subject knows what a reasonable person in the same situation would want to know before giving consent. *Consent* means an explicit agreement to participate. Fulfillment of informed consent, thus defined, can be challenging, especially when participants, for situational, intellectual, or cultural

reasons, have compromised autonomy, or do not accept the social norms that typically accompany social or behavioral research.

HISTORICAL BACKGROUND

After World War II, it was learned that Nazi scientists had used prisoners in brutal medical experiments. These crimes were investigated at the Nuremberg trials of Nazi war criminals and resulted in development of the Nuremberg Code, which requires that scientists obtain the informed consent of participants. In 1974, the passage of the U.S. National Research Act established the National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research (National Commission), and the use of Institutional Review Boards (IRBs) to oversee federally funded human research. The National Commission conducted hearings on ethical problems in human research, and it learned of biomedical research where concern for informed consent and well-being of subjects was overshadowed by other concerns. In their "Belmont Report," the National Commission formulated three principles to govern human research: respect for the autonomy and well-being of subjects, beneficence, and justice (see Ethical Principles). Informed consent is the main process through which respect is accorded to subjects.

ELEMENTS OF INFORMED CONSENT

The following elements of informed consent were set forth in the U.S. Federal Regulations of Human Research (Federal Regulations):

1. A statement that the study involves research, an explanation of the purposes of the research and expected duration of participation, a description of the procedures to be followed, and identification of any procedures that are experimental
2. A description of any reasonably foreseeable risks or discomforts to the subject
3. A description of any benefits to subjects or others that may reasonably be expected
4. A disclosure of appropriate alternative courses of treatment, if any, that might be advantageous to the subject (relevant primarily to clinical biomedical research)
5. A statement describing the extent, if any, to which confidentiality will be assured
6. For research involving more than minimal risk, a statement of whether any compensation or treatments are available if injury occurs
7. An explanation of whom to contact for answers to pertinent questions about the research, subjects' rights, and whom to contact in the event of a research-related injury
8. A statement that participation is voluntary, refusal to participate will involve no penalty, and the subject may discontinue participation at any time with impunity

Under the Federal Regulations, some or all of these elements of informed consent may be waived by the IRB if the research could not practicably be carried out otherwise. However, in waiving or changing the requirements, the IRB must document that there is no more than minimal risk to subjects; that the waiver will not adversely affect the rights and welfare of the subjects; and, whenever appropriate, that the subjects will be provided additional pertinent information after participation. These elements are often expressed in written form in a fairly standard format and are signed by research participants to indicate that they have read, understood, and agreed to those conditions. The requirement that the information be written or that the consent form be signed may be waived if the research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside of the research context. It may also be waived in order to preserve confidentiality if breach of confidentiality could harm the subject and if the signed consent form is the only record linking the subject and the research.

Misapplication of Informed Consent

Many investigators and IRBs operate as though a signed consent form fulfills ethical requirements of informed consent. Some consent forms are long, complex, inappropriate to the culture of the potential subject, and virtually impossible for most people to comprehend. Confusing informed consent with a signed consent form violates the ethical intent of informed consent, which is to communicate clearly and respectfully, to foster comprehension and good decision making, and to ensure that participation is voluntary.

Why do researchers and IRBs employ poor consent procedures? The Federal Regulations of Human

Research (45 CFR 4) were written to regulate human research sponsored by the Department of Health and Human Services (HHS). Most HHS-sponsored research is biomedical, and not surprisingly, the federal regulations governing human research were written with biomedical research in mind. This biomedical focus has always posed problems for social and behavioral research. These problems became more severe when the Office for Human Research Protection (OHRP) began responding to complaints about biomedical research by suspending the funded research of the entire institution and holding its IRB responsible for causing harm to human subjects. Responsibility was established by inspecting the paperwork of the IRB at the offending institution and noting inadequate paperwork practices. Such highly publicized and costly sanctions against research institutions resulted in an environment of fear in which IRBs learned that it is in their self-interest and that of their institution to employ highly bureaucratic procedures, emphasizing paperwork instead of effective communication. Results of this new regulatory climate include many bureaucratic iterations before a protocol is approved, and requirement of subjects' signatures on long, incomprehensible consent forms.

These problems are further compounded because Subpart A of 45 CFR 46 has been incorporated into the regulatory structure of 17 federal agencies (Housing & Urban Development, Justice, Transportation, Veterans Affairs, Consumer Product Safety, Environmental Protection, International Development, National Aeronautics & Space Administration, National Science Foundation, Agriculture, Commerce, Defense, Education, Energy, Health & Human Services, Social Security, and Central Intelligence Agency). Subpart A, now known as the Common Rule, governs all social and behavioral research at institutions that accept federal funding. Consequently, most IRBs, in their efforts to "go by the book," require a signed consent form even when this would be highly disrespectful (e.g., within cultures where people are illiterate or do not believe in signing documents). Within mainstream Western culture, Singer (1978) and Trice (1987) have found that a significant number of subjects refuse to participate in surveys, studies, or experiments if required to sign a consent form, but would gladly participate otherwise. Among subjects who willingly sign documents, most sign the consent form without reading it.

These problems can be avoided if investigators and IRBs remember to exercise their prerogative to exempt

specified categories of research, to conduct expedited review, or to waive written documentation of consent as permitted under 45 CFR 46.117(c), or if they recognize other ways of documenting informed consent (e.g., by having it witnessed, by the researcher's certification that the consent was conducted, or by audio or video recording of the consent procedure). Of most interest is paragraph (2) of 46.117(c) of the Federal Regulations: The IRB may waive the requirement for a signed consent form "if the research presents no more than minimal risk and involves no procedures for which written consent is normally required outside of the research context." Most social and behavioral research presents no more than minimal risk. The statement "no procedures for which written consent is normally required outside of the research context" refers to medical procedures. Consent forms are not used to document the patient's willingness to be interviewed or to complete questionnaires about their medical histories. Therefore, similarly personal questions might be asked of survey research respondents without triggering the requirement of documented informed consent, under the Common Rule. The following recommendations are offered in pursuit of intelligent interpretation of the Common Rule:

Informed consent should take the form of friendly, easily understood communication. When written, it should be brief and simply phrased at such a reading level that even the least literate subject can understand it. Subjects should have a comfortable context in which to think about what they have been told and to ask any questions that occur to them. The consent process may take many forms. For example, if conducting a housing study of elderly people at a retirement community, the interviewers should be mindful of the safety concerns of elders and of possible memory, auditory, and visual impairment. They might begin by holding a meeting at the community's meeting center to explain the purpose of the interviews and to answer any questions. They might arrange with the management to summarize the discussion in the community's newsletter, with photos of the interviewers. In preparation for the interviews, they might make initial contact with individual households by phone. Upon arriving, they might bring along the article so that residents can see that they are admitting bona fide interviewers. The interviewers might give a brief verbal summary of their purpose, confidentiality measures, and other pertinent elements of consent; obtain verbal consent; and provide a written

summary of the information for the subjects to keep as appropriate.

If there is some risk of harm or inconvenience to subjects, they should receive enough information to judge whether the risk is at a level they can accept. An adequate informed consent process can sort out those who would gladly participate from those who wish to opt out. There are risks or inconveniences that some will gladly undertake if given a choice, but that they would not want to have imposed upon them unilaterally. Others may decide that the experience would be too stressful, risky, or unpleasant for them.

The researcher should make it clear that the subject is free to ask questions at any time. Subjects do not necessarily develop questions or concerns about their participation until they are well into the research experience. For example, a discussion of confidentiality may not capture their attention until they are asked personal questions.

If subjects can refuse to participate by hanging up the phone or tossing a mailed survey away, the informed consent can be extremely brief. Courtesy and professionalism would require that the identity of the researcher and research institution be mentioned, along with the nature and purpose of the research. However, if there are no risks, benefits, or confidentiality issues involved, these topics and the right to refuse to participate need not be mentioned, when such details would be gratuitous. Detailed recitation of irrelevant information detracts from the communication. Assurances that there are no risks and descriptions of measures taken to ensure confidentiality are irrelevant, irritating, and have been found to reduce response rates. (See PRIVACY AND CONFIDENTIALITY.)

The cultural norms and lifestyle of subjects should dictate how to approach informed consent. It is disrespectful and foolhardy to seek to treat people in ways that are incompatible with their culture and circumstances. For example, a survey of homeless injection drug users should probably be preceded by weeks of “hanging out” and talking with them, using their informal communication “grapevine” to discuss their concerns, answer their questions, and negotiate the conditions under which the survey is conducted. Similar communication processes probably should precede research on runaways and drug-abusing street youngsters of indeterminate age. Such youngsters may be considered emancipated, or alternatively, it is

likely impossible to obtain parental consent because of unavailability of the parent. People who are functionally illiterate, suspicious of people who proffer documents or require signatures, and people from traditional cultures should also be approached in the style that is most comfortable to them. In community-based research, the formal or informal leaders of the community are the gatekeepers. The wise researcher consults with them to learn the concerns and perceptions of prospective subjects, how best to recruit, and how to formulate and conduct the informed consent process. Protocols for research on such populations should show evidence that the researcher is familiar with the culture of the intended research population and has arranged the informed consent and other research procedures accordingly.

Researchers and IRBs should flexibly consider a wide range of media for administering informed consent. Videotapes, brochures, group discussions, Web sites, and so on can be more appropriate ways of communicating with potential subjects than the kinds of legalistic formal consent forms that have often been used.

When written or signed consent places subjects at risk and the signed consent is the only link between the subject and the research, a waiver of informed consent should be considered. One useful procedure is to have a trusted colleague witness the verbal consent. Witnessed consent, signed by the witness, can also be used in the case of illiterate subjects.

In organizational research, both the superior who allows the study and the subordinates who are studied must consent. Both must have veto power. Superiors should not be in a position to coerce participation and should not be privy to information on whether subordinates consented, or to their responses.

Sometimes, people are not in a position to decide whether to consent until after their participation. For example, in brief sidewalk interviews, most respondents do not want to endure a consent procedure until they have participated. After the interview, they can indicate whether they will allow the interview to be used as research data in the study that is then explained to them. A normal consent procedure can ensue.

In summary, informed consent is a respectful communication process, not a consent form. By keeping this definition foremost in mind, the researcher can

decide on the appropriate content and format in relation to the context of the research.

Thanks are due Dr. Robert J. Levine for his critical comments of an earlier draft of some of this material.

—Joan E. Sieber

REFERENCES

- Fisher, C. B., & Wallace, S. A. (2000). Through the community looking glass: Reevaluating the ethical and policy implications of research on adolescent risk and psychopathology. *Ethics & Behavior*, 10(2), 99–118.
- Krueger, R. A. (1994). *Focus groups: A practical guide for applied research* (2nd ed.). Thousand Oaks, CA: Sage.
- Levine, E. K. (1982). Old people are not all alike: Social class, ethnicity/race, and sex are bases for important differences. In J. E. Sieber (Ed.), *The ethics of social research: Surveys and experiments* (pp. 127–144). New York: Springer-Verlag.
- Marshall, P. A. (1992). Research ethics in applied anthropology. *IRB: A Review of Human Subjects Research*, 14(6), 1–5.
- Melton, G. B., Levine, R. J., Koocher, G. P., Rosenthal, R., & Thompson, W. C. (1988). Community consultation in socially sensitive research: Lessons from clinical trials on treatments for AIDS. *American Psychologist*, 43, 573–581.
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report*. Washington, DC: U.S. Government Printing Office.
- Sieber, J. E. (Ed.). (1991). *Sharing social science data: Advantages and challenges*. Newbury Park, CA: Sage.
- Sieber, J. E. (1996). Typically unexamined communication processes in research. In B. H. Stanley, J. E. Sieber, & G. B. Melton (Eds.), *Research ethics: A psychological approach*. Lincoln: University of Nebraska Press.
- Singer, E. (1978). Informed consent: Consequences for response rate and response quality in social surveys. *American Sociological Review*, 43, 144–162.
- Singer, E., & Frankel, M. R. (1982). Informed consent procedures in telephone interviews. *American Sociological Review*, 47, 416–427.
- Stanton, A. L., Burker, E. J., & Kershaw, D. (1991). Effects of researcher follow-up of distressed subjects: Tradeoff between validity and ethical responsibility? *Ethics & Behavior*, 1(2), 105–112.
- Trice, A. D. (1987). Informed consent: VII: Biasing of sensitive self-report data by both consent and information. *Journal of Social Behavior and Personality*, 2, 369–374.
- U.S. Federal Regulations of Human Research 45 CFR 46. Retrieved from <http://ohrp.osophs.dhhs.gov>.

INSTRUMENTAL VARIABLE

One of the crucial assumptions in ORDINARY LEAST SQUARES (OLS) estimation is that the disturbance term

is not correlated with any of the explanatory variables. The rationale for this assumption is that we assume the explanatory variables and the disturbance have separate effects on the dependent variable. However, this is not the case if they are correlated with one another. Because we cannot separate their respective effects on the dependent variable, estimates are BIASED and inconsistent. Suppose we have a simple linear regression such as $Y = \beta_0 + \beta_1 X + \varepsilon$, where X and ε are positively correlated. As X moves up or down, so does ε , with direction depending on the nature of the correlation. Here, variations in Y associated with variations in X will also be related to those between Y and ε . This implies that we cannot estimate the population coefficients of X and ε precisely. It can be shown that this problem gets worse as we increase the sample size toward infinity.

There are three major cases where correlation between an explanatory variable and the disturbance is likely to occur: (a) The problem can occur when an explanatory variable is measured with error, making it stochastic. This is the so-called errors-in-variables problem. (b) The problem can occur when there is a lagged dependent variable accompanied by auto-correlated disturbances. (c) The problem can occur when there is a simultaneity problem. That is, the dependent variable contemporaneously causes the independent variable, rather than vice versa.

One approach to addressing this problem is to use an instrumental variable in place of the variable that is correlated with the disturbance. An *instrumental variable* is a proxy for the original variable that is not correlated with the disturbance term. In the simple linear regression equation above, an instrumental variable for X is one that is correlated with X (preferably highly) but not with ε . Let's define such a variable, Z , so that $X = Z + v$, where v is a random disturbance with zero mean. If we put this into the original equation above, then we will have $Y = \beta_0 + \beta_1(Z + v) + \varepsilon = \beta_0 + \beta_1 Z + \beta_1 v + \varepsilon = \beta_0 + \beta_1 Z + \varepsilon^*$, where $\varepsilon^* = \beta_1 v + \varepsilon$. Compared to the original equation, in which X and ε are correlated, this is not the case with Z and ε^* in the new equation. This is because ε^* contains $\beta_1 v$, which exists independently of Z . As we increase our sample infinitely, the correlation between Z and ε_i^* will approach zero. With such a proxy variable Z in hand, we still may not have an unbiased estimator for the coefficient of X , because a proxy is not equal to the true variable. However, as we increase the sample size, the estimator based upon Z will approach the true population coefficient.

A practical and serious problem with the instrumental variable approach is how to find a satisfactory proxy for an explanatory variable. If we have strong theory, this can guide us. If we don't, however, we have to rely on technical procedures to find it. For example, in TWO-STAGE LEAST SQUARES estimation, we regard as proxies of ENDOGENOUS VARIABLES the predicted values of the endogenous variables in the regressions on all EXOGENOUS VARIABLES. However, there is no guarantee that this kind of atheoretical procedure will provide a good proxy. As a general rule, the proxy variable should be strongly correlated with the original variable. Otherwise, estimates will be very imprecise.

—B. Dan Wood and Sung Ho Park

REFERENCES

- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.
- Kennedy, P. (1998). *A guide to econometrics* (4th ed.). Cambridge: MIT Press.

INTERACTION. See INTERACTION EFFECT; STATISTICAL INTERACTION

INTERACTION EFFECT

Interaction in statistics refers to the interplay among two or more INDEPENDENT VARIABLES as they influence a DEPENDENT VARIABLE. Interactions can be conceptualized in two ways. First is that the effect of each independent variable on an outcome is a conditional effect (i.e., depends on or is *conditional* on the value of another variable or variables). Second is that the *combined effect* of two or more independent variables is different from the sum of their individual effects; the effects are NONADDITIVE. Interactions between variables are termed HIGHER ORDER effects in contrast with the FIRST-ORDER effect of each variable taken separately.

As an example, consider people's intention to protect themselves against the exposure to the sun as a function of the distinct region of the country in which they reside (rainy northwest, desert southwest) and

Table 1 No Interaction Between Region and Risk—Arithmetic Mean Level of Intention to Protect Oneself Against the Sun as a Function of Region of the Country (Northwest, Southwest) and Objective Risk for Skin Cancer

		<i>Region of the Country</i>	
		<i>Northwest</i>	<i>Southwest</i>
<i>Risk of Skin Cancer</i>	<i>Low</i>	6	8
	<i>High</i>	7	9

their high versus low objective risk for skin cancer (see Table 1).

In Table 1, risk and region do not interact. In the northwest, those at low risk have a mean intention of 6; those at high risk have a mean intention of 7, or a 1-point difference in intention produced by risk. In the southwest, high risk again raises intention by 1 point, from 8 to 9. Looked at another way, people residing in the southwest have 2-point higher intention than in the northwest whether they are at low risk (8 vs. 6, respectively) or at high risk (9 vs. 7, respectively). The impact of risk is not conditional on region and vice versa. Finally, risk (low to high) raises intention by 1 point and region (northwest to southwest) raises intention by 2 points; we expect a $1 + 2 = 3$ -point increase in intention from low-risk individuals residing in the northwest to high-risk individuals residing in the southwest, and this is what we observe (from 6 to 9, respectively), an *additive* effect of risk and region.

In Table 2, risk and region interact. Low- versus high-risk individuals in the northwest differ by 1 point in intention, as in Table 1 (termed the *simple effect* of risk at one value of region, here northwest). However, in the southwest, the difference due to risk is quadrupled (8 vs. 12 points). Looked at another way, among low-risk individuals, there is only a 2-point difference in intention (6 to 8) as a function of region; for high-risk individuals, there is a 5-point difference (7 to 12). The impact of region is *conditional on* risk and vice versa. Instead of the additive 3-point effect of high risk and southwest region from Table 1, in Table 2, high-risk individuals in the southwest have a 6-point higher intention to sun protect than low-risk individuals in the northwest (12 vs. 6, respectively). The combined effects of region and risk are nonadditive, following the second definition of interaction.

Table 2 illustrates an interaction in terms of means of variables that are either CATEGORICAL (region) or

Table 2 Interaction Between Region and Risk—Arithmetic Mean Level of Intention to Protect Oneself Against the Sun as a Function of Region of the Country (Northwest, Southwest) and Objective Risk for Skin Cancer

		Region of the Country	
		Northwest	Southwest
Risk of Skin Cancer	Low	6	8
	High	7	12

have been artificially broken into categories (risk), as in ANALYSIS OF VARIANCE. Two CONTINUOUS VARIABLES may interact, or a continuous and a categorical variable may interact, as in MULTIPLE REGRESSION ANALYSIS. Figure 1 shows the interaction between region (categorical) and risk (with people sampled across a 6-point risk scale). Separate regression lines are shown for the northwest and southwest (termed *simple regression lines*). In the southwest (solid line), intention increases dramatically as risk increases. However, in the northwest (dashed line), risk has essentially no impact on intention. Region is a MODERATING VARIABLE or *moderator*, a variable that changes the relationship of another variable (here, risk) to the outcome. Put another way, the impact of risk is conditional on region. The vertical distance between the two lines shows the impact of region on intention. When risk is low (at 1 on the risk scale), the lines are close together; regional difference in intention is small. When risk is high (at 6 on the risk scale), regional difference is large. The impact of region is conditional on risk. Finally, being both at high risk and living in the southwest produces unexpectedly high intention, consistent with our second definition of nonadditive effect.

TYPES OF INTERACTIONS

The interaction between region and risk is *synergistic*: the effects of the variables together are greater than their sum. Being in a region of the country with intense sun (the southwest) and being at high personal risk combine to produce intention higher than the simple sum of their effects, a *synergistic interaction*. Alternatively, an interaction may be *buffering*, in that one variable may weaken the effect of another variable. Stress as an independent variable increases the risk of negative mental health outcomes. Having social support from family and friends weakens (or buffers)

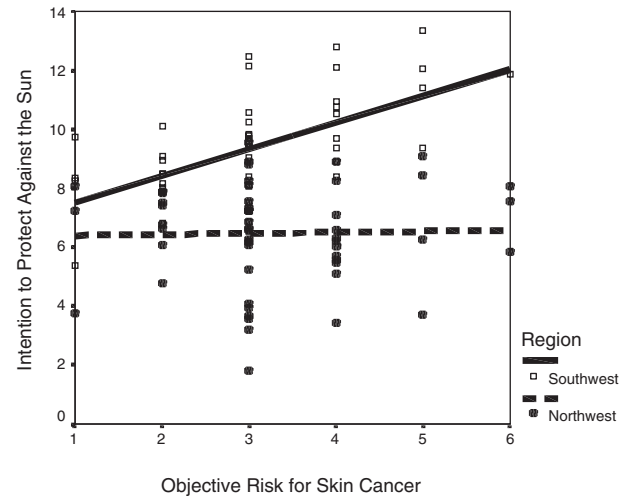


Figure 1 Interaction Between Region of the Country and Objective Risk for Skin Cancer in Prediction of Intention to Protect Against the Sun

NOTE: Region moderates the relationship of risk to intention. In the northwest, there is no effect of objective risk on intention, whereas in the southwest, there is a strong effect of objective risk on intention.

the effect of high stress on mental health outcomes, a *buffering interaction*.

The interaction illustrated in Figure 1 is an ORDINAL INTERACTION; that is, the order of levels of one variable is maintained across the range of the other variable. Here, intention is always higher in the southwest across the range of risk, characteristic of an ordinal interaction. An interaction may be *disordinal* (i.e., exhibit DISORDINALITY). Say that among high-risk individuals, those in the southwest had higher intention than those in the northwest. Among low-risk individuals, however, those in the northwest had higher intention. The interaction would be disordinal. This would be a *cross-over interaction*: Which value of one variable was higher would depend on the value of the other variable.

HISTORICAL DEVELOPMENT

In the 1950s through the 1970s, there was great emphasis on the characterization and interpretation of interactions between categorical variables (factors in the analysis of variance). Cohen (1978) turned attention to interactions between continuous variables in multiple linear regression. A decade later, Jaccard, Turrisi, and Wan (1990) and Aiken and West (1991)

extended the methodology for the post hoc probing and interpretation of continuous variable interactions. At present, there is focus on extending the characterization of interactions to other forms of regression analysis (e.g., LOGISTIC REGRESSION) and to interactions among latent variables in STRUCTURAL EQUATION MODELING.

—Leona S. Aiken

REFERENCES

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Cohen, J. (1978). Partialled products are interactions; partialled powers are curve components. *Psychological Bulletin*, 85, 858–866.
- Jaccard, J., Turrissi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. Newbury Park, CA: Sage.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

INTERCEPT

The intercept is defined as the value of a variable, Y , when a line or curve cuts through the y -axis. In a two-dimensional graph, the intercept is therefore the value of Y when X is zero. The units of the intercept are the same as the units of Y .

There are problems with the interpretation of intercepts in REGRESSION contexts: The zero value of the X variable often lies outside of the range of X observations, so the value of Y when X is zero is therefore not of interest in practical or policy contexts.

However, the intercept is a critical component of regression equations.

$$Y = a + bX$$

or

$$Y_i = \sum_j b_j X_{ij} + \varepsilon_i.$$

The subscript j indicates the different variables, starting with a neutral constant term b_0 . The subscript i indicates the cases in the sample data set. The intercept a in the simple bivariate equation corresponds to the constant term b_0 in the multivariate equation. The term b_0 is a multiplicative constant applied to the first variable, X_0 , whose values are all set equal to 1. Thus, a variable $X_0 = 1$ is created so that an intercept will be assigned to the whole equation.

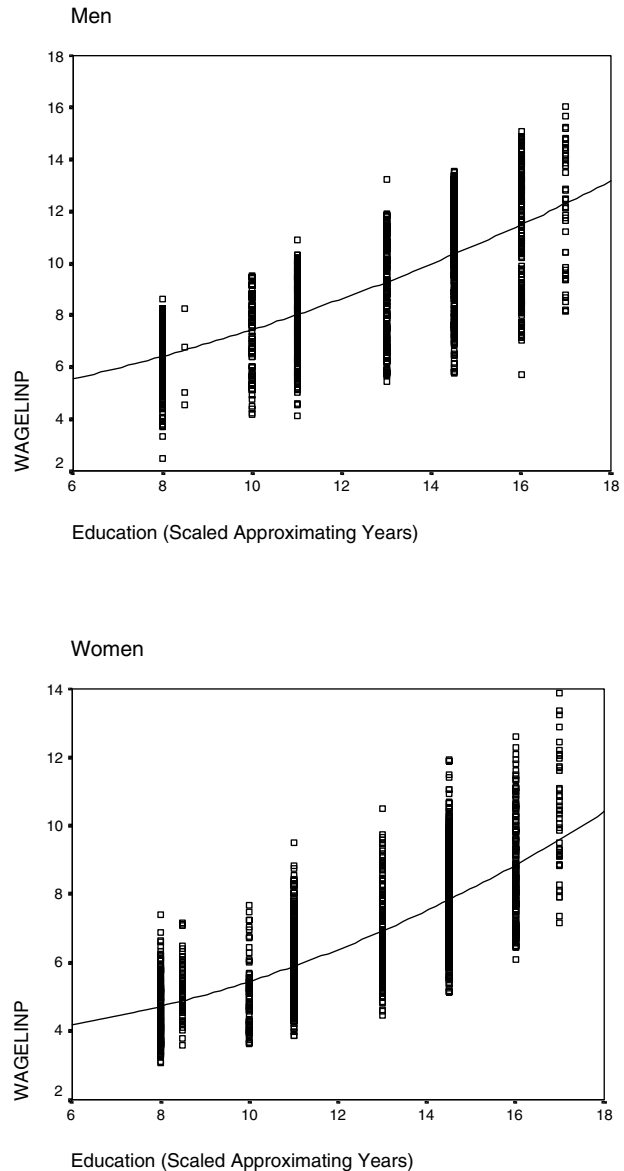


Figure 1 Hourly Wage Predicted by Regression: Education Dummy Variables

SOURCE: British Household Panel Survey, Wave 9, 1999–2000 (ESRC, 2001; Taylor, 2000).

Diagrams can illustrate for both linear and nonlinear equations where the intercept will fall. Some curves have no intersection with the x -axis, and these have no intercept. However, for linear regression, there is always a y -intercept unless the slope is infinite. The y -intercept can be set to zero if desired before producing an estimate.

An example illustrates how the intercept can be of policy interest even while remaining outside the range of reasonable values for the independent variable.

Regression equations for men's and women's wages per hour have been subjected to close scrutiny in view of the gender pay gap that exists in many countries (Monk-Turner & Turner, 2001). Other pay gaps, such as that between a dominant and a minority ethnic group, also get close scrutiny. A wage equation that includes dummy values for the highest level of education shows the response of wages to education among men and women using U.K. 2000 data (Walby & Olsen, 2002). The log of wages is used to represent the skewed wage variable, and linear regression is then used. The predictions shown below are based upon a series of dummy variables for education producing a spline estimate. The predictions are then graphed against years of formal education. The curve shown is that which best fits the spline estimates.

In mathematical terms, the true intercept occurs when $x = 0$, as if 0 years of education were possible. However, the intercept visible in Figure 1 (£5.75 for men and £4.25 for women) forms part of the argument about gender wage gaps: The women's line has a lower intercept than the men's line. Much debate has occurred regarding the intercept difference (Madden, 2000).

In REGRESSION analysis, the constant term in these equations is a factor that depends upon all 32 independent variables in the equation (for details, see Taylor, 2000; Walby & Olsen, 2002). However, the regression constant B_0 is not that which would be seen in this particular $x - y$ graph. B_0 is the constant in the multi-dimensional space of all the x_i s. However, a single graph showing one x_i and the predicted values of y has a visible intercept on the x_i -axis. Assumptions must be made regarding the implicit values of the other x_i s; they are usually held at their mean values, and these influence the intercept visible in the diagram of $x - \hat{y}$ (i.e., the predictions of y along the domain of x_i). Two-dimensional diagrams, with their visible intercepts and slopes, bear a complex relation to the mathematics of a complex MULTIPLE REGRESSION.

—Wendy K. Olsen

REFERENCES

- Economic and Social Research Council Research Centre on Micro-Social Change. (2001, February 28). *British Household Panel Survey* [computer file] (Study Number 4340). Colchester, UK: The Data Archive [distributor].
- Madden, D. (2000). Towards a broader explanation of male-female wage differences. *Applied Economics Letters*, 7, 765–770.
- Monk-Turner, E., & Turner, C. G. (2001). Sex differentials in the South Korean labour market. *Feminist Economics*, 7(1), 63–78.
- Taylor, M. F., with Brice, J., Buck, N., & Prentice-Lane, E. (Eds.). (2000). *British Household Panel Survey User Manual Volume B9: Codebook*. Colchester, UK: University of Essex.
- Walby, S., & Olsen, W. (2002, November). *The impact of women's position in the labour market on pay and implications for UK productivity*. Cabinet Office Women and Equality Unit. Retrieved from www.womenandequalityunit.gov.uk.

INTERNAL RELIABILITY

Most measures employed in the social sciences are constructed based on a small set of items sampled from the large domain of possible items that attempts to measure the same CONSTRUCT. It would be difficult to assess whether the responses to specific items on a measure can be generalized to the population of items for a construct if there is a lack of consistent responses toward the items (or a subset of the items) on the measure. In other words, high internal consistency suggests a high degree of interrelatedness among the responses to the items. Internal reliability provides an estimate of the consistency of the responses toward the items or a subset of the items on the measure.

There are a variety of internal reliability estimates, including CRONBACH'S ALPHA, Kuder Richardson 20 (KR20), SPLIT-HALF RELIABILITY, stratified alpha, maximal reliability, Raju coefficient, Kristof's coefficient, Angoff-Feldt coefficient, Feldt's coefficient, Feldt-Gilmer coefficient, Guttman λ_2 , maximized λ_4 , and Hoyt's analysis of variance. These estimates can all be computed after a single form of a measure is administered to a group of participants. In the remaining section, we will discuss only KR20, split-half reliability, stratified alpha, and maximal reliability.

When a measure contains dichotomous response options (e.g., yes/no, true/false, etc.), Cronbach's alpha takes on a special form, KR20, which is expressed as

$$\left[\frac{n}{n-1} \right] \left[\frac{\sigma_X^2 - \sum p_i(1-p_i)}{\sigma_X^2} \right],$$

where p_i is the proportion of the population endorsing item i as yes or true, and σ_X^2 is the variance of the observed score for a test with n components.

Split-half reliability, a widely used reliability estimate, is defined as the CORRELATION between two

halves of a test (r_{12}), corrected for the full length of the test. According to Spearman-Brown Prophecy formula, split-half reliability can be obtained by $\frac{2r_{12}}{1+r_{12}}$. It can also be calculated by

$$\frac{4r_{12} \times \sigma_1 \times \sigma_2}{\sigma_X^2},$$

which takes VARIANCES into consideration, where σ_1 and σ_2 are the STANDARD DEVIATIONS of the two halves of a test, and σ_X^2 is the variance of the entire test. Cronbach's alpha is the same as the average split-half reliability when split-half reliability is defined by the latter formula. However, if split-half reliability is calculated based on the Spearman-Brown Prophecy formula, Cronbach's alpha is equal to the average split-half reliability *only* when all item variances are the same.

If a test X is not unidimensional and is grouped into i subsets because of heterogeneous contents, stratified alpha and maximal reliability are better choices than Cronbach's alpha and its related estimates. Stratified alpha can be calculated as

$$1 - \sum \frac{\sigma_i^2(1 - \alpha_i)}{\sigma_X^2},$$

where σ_i^2 is the variance of the i th subset and α_i is the Cronbach's alpha of the i th subset. The maximal reliability estimate is appropriate when all items within a subset are parallel, although the items in different subsets may have different reliabilities and variances (Osburn, 2000). Maximal reliability can be derived by

$$R_I = \frac{A}{\frac{I}{[1+(I-1)\rho]} + A},$$

where I = number of subsets, ρ is the common correlation among the subsets

$$A = \frac{n_1\rho_1}{(1 - \rho_1)} + \frac{n_2\rho_2}{(1 - \rho_2)} + \cdots + \frac{n_i\rho_i}{(1 - \rho_i)}, n_i$$

is the number of items in the i th subset, and ρ_i is the reliability of the i th subset.

In general, the alternate-forms reliability is a preferred reliability estimate when compared with internal reliability or TEST-RETEST RELIABILITY. Unfortunately, the required conditions to estimate the alternate-forms reliability are not easily met in practice (see RELIABILITY). Compared to test-retest reliability, internal reliability is flexible and tends not to be affected by problems such as practice, fatigue, memory, and motivation. It is flexible because it requires only a

single administration and can be applied to a variety of circumstances, including a test containing multiple items, a test containing multiple subscales, or a score containing multiple raters' judgments. Internal consistency reliability is generally not an appropriate reliability estimate when (a) the test is heterogeneous in content, (b) a battery score is derived from two or more heterogeneous tests, or (c) the test is characterized as a speeded test.

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.
- Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, 5, 343–355.

INTERNAL VALIDITY

Internal validity refers to the confidence with which researchers can make *causal* inferences from the results of a particular empirical study. In their influential paper on “Experimental and Quasi-Experimental Designs for Research,” Campbell and Stanley (1963) characterized internal validity as the *sine qua non* of experimental research. This is because the very purpose of conducting an experiment is to test causal relationships between an INDEPENDENT VARIABLE and a DEPENDENT VARIABLE. Hence, it is appropriate that the primary criterion for evaluating the results of an experiment is whether valid causal conclusions can be drawn from its results. However, the concept of internal validity can also be applied to findings from correlational research anytime that causal inferences about the relationship between two variables are being drawn.

It is important to clarify that internal validity does not mean that a particular independent variable is the *only* cause of variation in the dependent variable—only that it has some independent causal role. For instance, variations among individuals in their attitudes toward a social program can be caused by many factors, including variations in personal history, self-interest, ideology, and so on. But if the mean attitude of a group of people that has been exposed to a particular persuasive message is more favorable than that of a

group of people that did not receive the message, the question of interest is whether the message played a role in causing that difference. We can draw such a conclusion only if no other relevant causal factors were correlated with whether an individual received the message or not. In other words, there are no plausible alternative explanations for the covariation between attitude and receipt of the message. The existence of correlated rival factors is usually referred to as a **CONFOUNDING** of the experimental treatment because the potential effects of the variable under investigation cannot be separated from the effects of these other potential causal variables.

In this example, the most obvious challenge to causal inference would be any kind of self-selection to message conditions. If participants in the research were free to decide whether or not they would receive the persuasive message, then it is very likely that those who already had more favorable attitudes would end up in the message-receipt group. In that case, the difference in attitudes between the groups after the message was given may have been predetermined and had nothing to do with the message itself. This does not necessarily mean that the message did not have a causal influence, only that we have a reasonable alternative that makes this conclusion uncertain. It is for this reason that **RANDOM ASSIGNMENT** is a criterial feature of good experimental design. Assigning participants randomly to different levels of the independent variable rules out self-selection as a threat to internal validity.

BASIC THREATS TO INTERNAL VALIDITY

A research study loses internal validity when there is reason to believe that obtained differences in the dependent variable would have occurred even if exposure to the independent variable had not been manipulated. In addition to self-selection as a potential confounding factor, Campbell and Stanley (1963) described several other generic classes of possible threats to internal validity. These are factors that could be responsible for variation in the dependent variable, but they constitute threats to internal validity only if the research is conducted in such a way that variations in these extraneous factors become correlated with variation in the independent variable of interest. The types of potential confounds discussed by Campbell and Stanley (1963) included the following:

History. Differences in outcomes on the dependent variable measured at two different times may

result from events other than the experimental variable that have occurred during the passage of time between measures. History is potentially problematic if the dependent variable is measured before and after exposure to the independent variable, where other intervening events could be a source of change.

Maturation. Another class of effects that may occur over the passage of time between measures on the dependent variable involves changes in the internal conditions of the participants in the study, such as growing older, becoming more tired, less interested, and so on. (These are termed *maturation effects* even though some representatives of this class, such as growing tired, are not typically thought of as being related to physical maturation.)

Testing. Participants' scores on a second administration of the dependent variable may be affected by the fact of their having been exposed to the measure previously. Thus, testing itself could be a source of change in the dependent variable. This would also be a problem if some individuals in a study had received prior testing and others had not.

Instrumentation. Changes across time in dependent variable scores may be caused by changes in the nature of the measurement "instrument" (e.g., changes in attitudes of observers, increased sloppiness on the part of test scorers, etc.) rather than by changes in the participants being measured.

Statistical Regression. Unreliability, or error of measurement, will produce changes in scores on different measurement occasions, and these scores are subject to misinterpretation if participants are selected on the basis of extreme scores at their initial measurement session.

Experimental Mortality. If groups are being compared, any selection procedures or treatment differences that result in different proportions of participants dropping out of the experiment may account for any differences obtained between the groups in the final measurement.

Selection-History Interactions. If participants have been differentially selected for inclusion in comparison groups, these specially selected groups may experience differences in history, maturation, testing, and so on, which may produce differences in the final measurement on the dependent variable.

Again, it should be emphasized that the presence of any of these factors in a research study

undermines internal validity if and only if it is differentially associated with variations in the independent variable of interest. Events other than the independent variable may occur that influence the outcomes on the dependent variable, but these will not reduce internal validity if they are not systematically correlated with the independent variable.

—Marilynn B. Brewer

REFERENCES

- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.
- Crano, W. D., & Brewer, M. B. (2002). *Principles and methods of social research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

INTERNET SURVEYS

The term “Internet survey” refers simply to any SURVEY where the data are collected via the Internet. Internet surveys are having an impact on the survey profession like few other innovations before. In the short time that the Internet (and, more particularly, the World Wide Web) has been widely available, the number of Web survey software products, online panels, and survey research organizations offering Web survey services, either as a sole data collection mode or in combination with traditional methods of survey research, has proliferated. However, the reception within the profession has been mixed, with some hailing online surveys as the replacement for all other forms of survey data collection, and others viewing them as a threat to the future of the survey industry. Although much is still unknown about the value of Internet surveys, research on their utility is proceeding apace.

TYPES OF INTERNET SURVEYS

The Internet can be used for survey data collection in several different ways. A distinction can be made between those surveys that execute on a respondent’s machine (client-side) and those that execute on the survey organization’s Web server (server-side). Key client-side survey approaches include e-mail surveys

and downloadable executables. In each of these cases, the instrument is transmitted to sample persons, who then answer the survey questions by using the reply function in the e-mail software, entering responses using a word processor, or using software installed on their computers. Once complete, the answers are transmitted back to the survey organization. Server-side systems typically involve the sample person completing the survey while connected to the Internet through a browser, with the answers being transmitted to the server on a flow basis. Interactive features of the automated survey instrument are generated by scripts on the Web server. A key distinction between these two approaches is whether the Internet connection is “on” while the respondent is completing the survey. Web surveys are the prime example of the second type and are by far the dominant form of Internet survey prevalent today.

There are variations of these two basic types. For example, some surveys consist of a single scrollable HTML form, with no interaction with the server until the respondent has completed the survey and pressed the “submit” button to transmit the information. Similarly, client-side interactivity can be embedded in HTML forms using JavaScript, DHTML, or the like. These permit a variety of dynamic interactions to occur without the involvement of the Web server.

The typical Web survey involves one or more survey questions completed by a respondent on a browser connected to the World Wide Web. Sampling of persons may take many forms, ranging from self-selected samples to list-based approaches (see Couper, 2000, for a review). In the latter case, the invitation to participate is often sent via e-mail, with the survey URL embedded in the e-mail message, along with an ID and password to control access. Web surveys are also increasingly used in mixed-mode designs, where mail surveys are sent to respondents, with an option to complete the survey online.

ADVANTAGES OF INTERNET SURVEYS

The primary argument in favor of Internet surveys is cost. Although fixed costs (hardware and software) may be higher than mail, the per-unit cost for data collection is negligible. Mailing and printing costs are eliminated using fully automated methods, as are data keying and associated processing costs. These reduce both the cost of data collection and the time taken to obtain results from the survey.

Another benefit of Web surveys is their speed relative to other modes of data collection. It is often the case that a large proportion of all responses occur within 24 hours of sending the invitation.

Web surveys also offer the advantages associated with self-administration, including the elimination or reduction of interviewer effects, such as socially desirable responding. RESPONDENTS can answer at their own pace and have greater control over the process. At the same time, Web surveys offer all the advantages associated with computerization. These include the use of complex branching or routing, randomization, edit checks, and so on that are common in COMPUTER-ASSISTED INTERVIEWING. Finally, such surveys can exploit the rich, graphical nature of the Web by including color, images, and other multimedia elements into the instruments.

DISADVANTAGES OF INTERNET SURVEYS

The biggest drawbacks of Internet surveys relate to representational issues. Internet access is still far from universal, meaning that large portions of the POPULATION are not accessible via this method. To the extent that those who are covered by the technology differ from those who are not, coverage BIAS may be introduced in survey estimates. Of course, for some populations of interest (students, employees in high-tech industries, visitors to a Website, etc.), this is less of a concern. A related challenge is one of SAMPLE. No list of Internet users exists, and methods akin to RANDOM DIGIT DIALING (RDD) for telephone surveys are unlikely to be developed for the Internet, given the variety of formats of e-mail addresses.

NONRESPONSE is another concern. The response rates obtained in Internet surveys are typically lower than those for equivalent paper-based methods (see Fricker & Schonlau, 2002). In mixed-mode designs, where sample persons are offered a choice of paper or Web, they overwhelmingly choose the former.

Another potential drawback is the anonymity of the Internet. There is often little guarantee that the person to whom the invitation was sent actually completed the survey, and that they took the task seriously. Despite this concern, results from Internet surveys track well with other methods of survey data collection, suggesting that this is not a cause of great concern.

SUMMARY

Despite being a relatively new phenomenon, Internet surveys have spread rapidly, and there are already

many different ways to design and implement online surveys. Concerns about representation may limit their utility for studies of the general population, although they may be ideally suited for special population studies (among groups with high Internet penetration) and for those studies where representation is less important (e.g., experimental studies). Given this, Internet surveys are likely to continue their impact on the survey industry as cutting-edge solutions to some of the key challenges are found and as the opportunities offered by the technology are exploited.

—Mick P. Couper

REFERENCES

- Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64(4), 464–494.
- Fricker, R. D., & Schonlau, M. (2002). Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods*, 14(4), 347–365.

INTERPOLATION

Interpolation means to estimate a value of a FUNCTION or series between two known values. Unlike EXTRAPOLATION, which estimates a value outside a known range from values within a known range, interpolation is concerned with estimating a value inside a known range from values within that same range.

There are, broadly, two approaches to interpolation that have been used in the social sciences. One is mathematical, and the other is statistical. In mathematical interpolation, we estimate an unknown value by assuming a deterministic relationship between two known data points. For example, suppose we are trying to find the 99% critical value for a CHI-SQUARE variable with 11.3 degrees of freedom (e.g., see Greene, 2000, p. 94). Note that chi-square tables are given only for integer values. However, the actual chi-square distribution is also defined for noninteger values. We can interpolate to noninteger values for this example as follows. First, obtain the 99% critical values of the chi-square distribution with 11 and 12 degrees of freedom from the tables. Then, interpolate linearly between the reciprocals of the degrees of freedom. Thus, given that $\chi_{11}^2 \leq 0.99 = 24.725$ and $\chi_{12}^2 \leq 0.99 = 26.217$, we have

$$\begin{aligned} \chi_{11.3}^2 \leq 0.99 &= 26.217 + \frac{\frac{1}{11.3} - \frac{1}{12}}{\frac{1}{11} - \frac{1}{12}} (24.725 - 26.217) \\ &= 25.209. \end{aligned}$$

This example uses linear interpolation. However, there are other problems where interpolation might involve inserting unknown values with nonlinear functions such as a cosine, polynomial, cubic spline, or other functions. The method of mathematical interpolation chosen should be consistent with the desired result.

Statistical interpolation estimates unknown values between known values by relying on information about the dynamics or DISTRIBUTION of the whole data set. Notice here the difference from the mathematical approach. In the previous example, we just draw a linear or smooth line between two data points to get an unknown value, regardless of the specific internal structure of the data set. However, in the statistical method, we are concerned with the specific dynamics and distribution inherent in the data set and project this information to get the unknown value. We can secure this information by either finding a related data set or simply assuming an internal structure of the data. For example, suppose we have a variable *X* that consists of quarterly TIME-SERIES data, and we want to convert *X* into a monthly time series. One approach is to find *Y*, a monthly time series, that is highly correlated with the original *X* series. We would first eliminate the differences between the quarterly movements of the two data series and then project the monthly series of *Y* onto *X*. Different outcomes can be obtained depending on how we eliminate the quarterly differences between *X* and *Y* and how we project *Y* onto the *X* series. Suppose, however, we are unable to find a data series that is related to *X*; we can also simply assume an internal dynamic. For example, we might have quarterly data prior to some date, and monthly data after that date (e.g., the University of Michigan's Index of Consumer Sentiment before and after 1978). We could simply assume that the monthly data structure is consistent with the unknown monthly dynamics of the earlier quarterly data. Using this assumption, we could construct an ARIMA model for the *X* series, in which we specify a combination of AUTOREGRESSIVE, MOVING AVERAGE, and RANDOM WALK movements, and project this specification into the unknown monthly series of *X*.

There are many technically advanced methods of interpolation that have been derived from the two broad categories discussed above. However, it should be emphasized that none of them can eliminate the uncertainty inherent in interpolation when true values do not exist. This is because it is impossible to find the true value of unknown data. In some cases, interpolation

even might end up with a serious distortion of the true data structure. Thus, caution is required when creating and using interpolated data.

—B. Dan Wood and Sung Ho Park

REFERENCES

- Friedman, M. (1962). The interpolation of time series by related series. *Journal of the American Statistical Association*, 57, 729–757.
- Greene, W. H. (2000). *Econometric analysis* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

INTERPRETATIVE REPERTOIRE

Social psychological DISCOURSE ANALYSIS aims in part to identify the explanatory resources that speakers use in characterizing events, taking positions, and making arguments. *Interpretative repertoire* is a term used to define those resources with regard to a particular theme or topic. Typically, these themes or topics are sought in interview data on matters of ideological significance, such as race, gender, and power relations. Interpretative repertoires occupy a place in analysis that seeks to link the details of participants' descriptive practices to the broader ideological and historical formations in which those practices are situated.

The essence of interpretative repertoires is that there is always more than one of them available for characterizing any particular state of affairs. Usually, there are two, used in alternation or in opposition, but there may be more. Repertoires are recurrent verbal images, metaphors, figures of speech, and modes of explanation. There is no precise definition or heuristic for identifying them. Rather, the procedure is to collect topic-relevant discourse data and examine them for recurrent interpretative patterns. This resembles the stages of developing GROUNDED THEORY, but it is a feature of interpretative repertoire analysis that it remains closely tied to the details of the discourse data, informed by the procedures of DISCOURSE ANALYSIS, rather than resulting in an abstract set of codes and categories. Interpretative repertoires are best demonstrated by examples.

The first use of the term was by Nigel Gilbert and Michael Mulkay (1984) in a study of scientific discourse, where they discovered two contrasting and functionally different kinds of explanation. There was an *empiricist repertoire*—the impersonal,

method-based, data-driven account of findings and theory choice typically provided in experimental reports. There was also a *contingent repertoire*—an appeal to personal motives, insights, and biases; social settings; and commitments—a realm in which speculative guesses and intuitions can operate and where conclusions and theory choice may give rise to, rather than follow from, the empirical work that supports them. Whereas the empiricist repertoire was used in making factual claims and supporting favored theories, the contingent repertoire was used in accounting for how and when things go wrong, particularly in rival laboratories, and with regard to discredited findings.

Margaret Wetherell and Jonathan Potter (1992) conducted interviews with New Zealanders on matters of race and ethnicity. They identified, among others, a *heritage repertoire* in talk about culture and cultural differences. This consisted of descriptions, metaphors, figures of speech, and arguments that coalesced around the importance of preserving traditions, values, and rituals. An alternative *therapy repertoire* defined culture as something that estranged Maoris might need in order to be whole and healthy again. Rather than expressing different attitudes or beliefs across different speakers, each repertoire could be used by the same people, because each had its particular rhetorical uses. The therapy metaphor, for example, was useful in talk about crime and educational failure. Repertoires of this kind, although often contradictory, were typically treated by participants as unquestionable common sense each time they were used. This commonsense status of interpretative repertoires enables links to the study of ideology.

—Derek Edwards

REFERENCES

- Gilbert, G. N., & Mulkay, M. (1984). *Opening Pandora's box: A sociological analysis of scientists' discourse*. Cambridge, UK: Cambridge University Press.
- Wetherell, M., & Potter, J. (1992). *Mapping the language of racism: Discourse and the legitimization of exploitation*. Brighton, UK: Harvester Wheatsheaf.

INTERPRETIVE BIOGRAPHY

The interpretive biographical method involves the studied use and collection of personal life documents, stories, accounts, and narratives that describe

turning-point moments in individuals' lives (see Denzin, 1989, Chap. 2). The subject matter of the biographical method is the life experiences of a person. When written in the first person, it is called an autobiography, life story, or life history. When written by another person, observing the life in question, it is called a biography.

The following assumptions and arguments structure the use of the biographical method. Lives are only known through their representations, which include performances such as stories, personal narratives, or participation in ritual acts. LIVED EXPERIENCE and its representations are the proper subject matter of sociology. Sociologists must learn how to connect and join biographically meaningful experiences to society-at-hand, and to the larger culture- and meaning-making institutions of the past postmodern period. The meanings of these experiences are given in the performances of the people who experience them. A preoccupation with method, with the VALIDITY, RELIABILITY, GENERALIZABILITY, and theoretical relevance of the biographical method, must be set aside in favor of a concern for meaning, performance, and interpretation. Students of the biographical method must learn how to use the strategies and techniques of performance theory, literary interpretation, and criticism. They must bring their use of the method in line with recent STRUCTURALIST and POSTSTRUCTURALIST developments in critical theory concerning the reading and writing of social texts. This will involve a concern with HERMENEUTICS, SEMIOTICS, CRITICAL RACE, FEMINIST, and postcolonial theory, as well as pedagogical cultural studies.

Lives and the biographical methods that construct them are literary productions. Lives are arbitrary constructions, constrained by the cultural writing practices of the time. These cultural practices lead to the inventions and influences of gendered, knowing others who can locate subjects within familiar social spaces, where lives have beginnings, turning points, and clearly defined endings. Such texts create "real" people about whom truthful statements are presumably made. In fact, these texts are narrative fictions, cut from the same kinds of cloth as the lives they tell about. When a writer writes a biography, he or she writes himself or herself into the life of the subject written about. When the reader reads a biographical text, that text is read through the life of the reader. Hence, writers and readers conspire to create the lives they write and read about. Along the way, the produced text is cluttered by

the traces of the life of the “real” person being written about.

These assumptions, or positions, turn on, and are structured by, the problem of how to locate and interpret the subject in biographical materials. All biographical studies presume a life that has been lived, or a life that can be studied, constructed, reconstructed, and written about. In the present context, a *life* refers to two phenomena: lived experiences, or conscious existence, and person. A person is a self-conscious being, as well as a named, cultural object or cultural creation. The consciousness of the person is simultaneously directed to an inner world of thought and experience and to an outer world of events and experience. These two worlds, the inner and the outer, are opposite sides of the same process, for there can be no firm dividing line between inner and outer experience. The biographical method recognizes this fact about human existence, for its hallmark is the joining and recording of these two structures of experience in a personal document, a performative text.

—Norman K. Denzin

REFERENCE

Denzin, N. K. (1989). *Interpretive biography*. Newbury Park, CA: Sage.

INTERPRETIVE INTERACTIONISM

Interpretive interactionists study personal troubles and turning-point moments in the lives of interacting individuals. Interpretive interactionism implements a critical, existential, interpretive approach to human experience and its representations. This perspective follows the lead of C. Wright Mills, who argued in *The Sociological Imagination* (1959) that scholars should examine how the private troubles of individuals, which occur within the immediate world of experience, are connected to public issues and to public responses to these troubles. Mills’ sociological imagination was biographical, interactional, and historical. Interpretive interactionism implements Mills’ project.

Interpretive interactionists attempt to make the problematic LIVED EXPERIENCE of ordinary people available to the reader. The interactionist interprets these worlds, their effects, meanings, and representations. The research methods of this approach include

performance texts; AUTOETHNOGRAPHY; poetry; fiction; open-ended, creative interviewing; document analysis; SEMIOTICS; LIFE HISTORY; life-story; personal experience and self-story construction; PARTICIPANT OBSERVATION; and THICK DESCRIPTION. The term *interpretive interactionism* signifies an attempt to join traditional SYMBOLIC INTERACTIONIST thought with critical forms of interpretive inquiry, including reflexive participant observation, literary ethnography, POETICS, life stories, and TESTIMONIOS.

Interpretive interactionism is not for everyone. It is based on a research philosophy that is counter to much of the traditional scientific research tradition in the social sciences. Only people drawn to the qualitative, interpretive approach are likely to use its methods and strategies. These strategies favor the use of minimal theory and value tightly edited texts that show, rather than tell.

Interpretive interactionists focus on those life experiences, which radically alter and shape the meanings people give to themselves and their life projects. This existential thrust leads to a focus on the “epiphany.” In epiphanies, personal character is manifested and made apparent. By studying these experiences in detail, the researcher is able to illuminate the moments of crisis that occur in a person’s life. These moments are often interpreted, both by the person and by others, as turning-point experiences. Having had this experience, the person is never again quite the same. Interpretive interactionists study these experiences, including their representations, effects, performances, and connections to broader historical, cultural, political, and economic forces.

—Norman K. Denzin

REFERENCES

- Denzin, N. K. (1989a). *Interpretive biography*. Newbury Park, CA: Sage.
 Denzin, N. K. (1989b). *Interpretive interactionism*. Newbury Park, CA: Sage.
 Denzin, N. K. (1997). *Interpretive ethnography*. Thousand Oaks, CA: Sage.
 Mills, C. W. (1959). *The sociological imagination*. New York: Oxford University Press.

INTERPRETIVISM

Interpretivism is a term used to identify approaches to social science that share particular ONTOLOGICAL

and EPISTEMOLOGICAL assumptions. The central tenet is that because there is a fundamental difference between the subject matters of the natural and social sciences, the methods of the natural sciences cannot be used in the social sciences. The study of social phenomena requires an understanding of the social worlds that people inhabit, which they have already interpreted by the meanings they produce and reproduce as a necessary part of their everyday activities together. Whereas the study of natural phenomena requires the scientist to interpret nature through the use of scientific concepts and theories, and to make choices about what is relevant to the problem under investigation, the social scientist studies phenomena that are already interpreted.

ORIGINS

Interpretivism has its roots in the German intellectual traditions of HERMENEUTICS and PHENOMENOLOGY, in particular in the work of Max Weber (1864–1920) and Alfred Schütz (1899–1959). Like Dilthey before them, they sought to establish an objective science of the subjective with the aim of producing verifiable knowledge of the meanings that constitute the social world. Their attention focused on the nature of meaningful social action, its role in understanding patterns in social life, and how this meaning can be assessed.

Rather than trying to establish the actual meaning that a social actor gives to a particular social action, Weber and Schütz considered it necessary to work at a higher level of generality. Social regularities can be understood, perhaps explained, by constructing models of typical meanings used by typical social actors engaged in typical courses of action in typical situations. Such models constitute tentative HYPOTHESES to be tested. Only social action that is rational in character, that is consciously selected as a means to some goal, is regarded as being understandable.

According to Weber (1964, p. 96), subjective meanings may be of three kinds: They may refer to the actual, intended meanings used by a social actor; the average or approximate meanings used by a number of social actors; or the typical meanings attributed to a hypothetical social actor.

Drawing on Dilthey, Weber distinguished between three modes of understanding: two broad types—*rational* understanding, which produces a clear intellectual grasp of social action in its context of meaning, and *empathetic* or appreciative understanding, which

involves grasping the emotional content of the action—and two versions of *rational* understanding, *direct* and *motivational* understanding. *Direct* understanding of a human expression or activity is like grasping the meaning of a sentence, a thought, or a mathematical formula. It is an immediate, unambiguous, matter-of-fact kind of understanding that occurs in everyday situations and that does not require knowledge of a wider context. *Motivational* understanding, on the other hand, is concerned with the choice of a means to achieve some goal.

Weber was primarily concerned with this motivational form of rational action. He regarded human action that lacks this rational character as being unintelligible. The statistical patterns produced by quantitative data, such as the relationship between educational attainment and occupational status, are not understandable on their own. Not only must the relevant actions that link the two components of the relationship be specified, but the meaning that is attached to these actions must also be identified.

Weber acknowledged that motives can be both rational and nonrational, they can be formulated to give action the character of being a means to some end, or they can be associated with emotional (affectual) states. Only in the case of rational motives did he consider that it was possible to formulate social scientific explanations.

Weber regarded meaningful interpretations as plausible hypotheses that need to be tested, and he recommended the use of COMPARATIVE RESEARCH to investigate them. However, if this is not possible, he suggested that the researcher will have to resort to an “imaginary experiment,” the thinking away of certain elements of a chain of motivation and working out the course of action that would probably then ensue.

Weber was not primarily concerned with the actor’s own subjective meaning but, rather, with the meaning of the situation for a constructed hypothetical actor. Neither was he particularly interested in the specific meanings social actors give to their actions but with approximations and abstractions. He was concerned with the effect on human behavior not simply of meaning but of meaningful social relations. His ultimate concern was with large-scale uniformities. Because he did not wish to confine himself to either contemporary or micro situations, he was forced to deal with the observer’s interpretations. Nevertheless, he regarded such interpretations only as hypotheses to be tested.

Weber's desire to link statistical uniformities with VERSTEHEN (interpretive understanding) has led to a variety of interpretations of his version of interpretivism. Positivists who have paid attention to his work have tended to regard the verstehen component as simply a potential source of hypotheses, whereas interpretivists have tended to ignore his concern with statistical uniformities and causal explanation.

Because Weber never pursued methodological issues beyond the requirements of his own substantive work, he operated with many tacit assumptions, and some of his concepts were not well developed. These aspects of Weber's work have been taken up sympathetically by Schütz (1976), for whose work Weber and Husserl provided the foundations. Schütz's main aim was to put Weber's sociology on a firm foundation. In pursuing this task, he offered a methodology of ideal types. Schütz argued that, in assuming that the concept of the meaningful act is the basic and irreducible component of social phenomena, Weber failed, among other things, to distinguish between the meaning the social actor *works with* while action is taking place, the meaning the social actor *attributes to* a completed act or to some future act, and the meaning the sociologist *attributes to* the action. In the first case, the meaning worked with during the act itself, and the context in which it occurs, is usually taken for granted. In the second case, the meaning attributed will be in terms of the social actor's goals. In the third case, the context of meaning will be that of the observer, not the social actor. Weber appeared to assume that the latter is an adequate basis for arriving at the social actor's attributed meaning and that there will be no disputes between actors, or between actors and observers, about the meaning of a completed or future act.

Schütz has provided the foundation for a methodological bridge between the meaning social actors attribute and the meaning the social scientists must attribute in order to produce an adequate theory. He argued that the concepts and meanings used in social theories must be derived from the concepts and meanings used by social actors. In Schütz's version of interpretivism, everyday reality is paramount; he argued that it is the social actor's, not the social investigator's, point of view that is the basis of any accounts of social life. The work of Weber and Schütz has provided the foundations for the major branches of interpretivism.

CRITICISMS

The critics of interpretivism have come from within as well as from without the approach. For example, Giddens (1984) has argued that it is misleading to imply that competent social actors engage in an ongoing monitoring of their conduct and are thus aware of both their intentions and the reasons for their actions. However, it is usually only when actors carry out retrospective inquiries into their own conduct, when others query their actions, or when action is disturbed or breaks down that this reflection occurs. Most of the time, action proceeds without reflective monitoring.

In response to the view of some interpretivists, that the social scientists must just report social actors' accounts of their actions and not interpret or theorize them, Rex (1974) and others have argued that social scientists should be able to give a different and competing account of social actors' actions from what those actors provide. Bhaskar (1979) has referred to this as the *linguistic fallacy*, which is based on a failure to recognize that there is more to reality than is expressed in the language of social actors. Bhaskar (1979) and Outhwaite (1987) have argued that interpretivism also commits the *epistemic fallacy* of assuming that our access to the social world is only through our understanding of interpretive processes—that this is all that exists. Other critics, such as Giddens and Rex, have argued that interpretivism fails to acknowledge the role of institutional structures, particularly divisions of interest and relations of power.

—Norman Blaikie

REFERENCES

- Bhaskar, R. (1979). *The possibilities of naturalism: A philosophical critique of the contemporary human sciences*. Brighton, UK: Harvester.
- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Cambridge, UK: Polity.
- Outhwaite, W. (1987). *New philosophies of social science: Realism, hermeneutics and critical theory*. London: Macmillan.
- Rex, J. (1974). *Sociology and the demystification of the modern world*. London: Routledge & Kegan Paul.
- Schütz, A. (1976). *The phenomenology of the social world* (G. Walsh & F. Lehnert, Trans.). London: Heinemann.
- Weber, M. (1964). *The theory of social and economic organization* (A. M. Henderson & T. Parsons, Trans.). New York: Free Press.

INTERQUARTILE RANGE

For quantitative data that have been sorted numerically from low to high, the interquartile range is that set of values between the 25th PERCENTILE (or P_{25}) in the set and the 75th percentile (P_{75}) in the set. Because this range contains half of all the values, it is descriptive of the “middle” values in a DISTRIBUTION. The range is referred to as the interquartile range because the 25th percentile is also known as the first QUARTILE and the 75th percentile as the third quartile.

Consider the ordered data set {3, 5, 7, 10, 11, 12, 14, 24}. There are eight values in this set. Eliminating the bottom two (which constitute the bottom 25%) and the top two (which constitute the top 25%) leaves the middle 50%. Their range is 7 to 12, which could be taken to be the interquartile range for the eight observations; however, because P_{25} must be a value below which 25% fall and above which 75% fall, it is conventional to define P_{25} as the value halfway between the boundary values, 5 and 7. This is 6. In like manner, the 75th percentile would be defined as 13, the point halfway between the boundary values 12 and 14. Thus, the interquartile range for the sample dataset is $13 - 6$, or 7.

Reporting interquartile ranges as part of data description enriches the description of central tendency that is provided by point values such as the MEDIAN (or second quartile) and the arithmetic MEAN. The greatest use of the interquartile range, however, is in the creation of BOXPLOTS and in the definition of OUTLIERS or “outside values,” which are data values that are quantitatively defined as highly atypical. A widely accepted definition of an outlier is a data value that lies more than 1.5 times the interquartile range above P_{75} or below P_{25} (Tukey, 1977, p. 44ff). In boxplots, these values are given a distinctive marking, typically an asterisk. For the data set above, 24 is an outlier because it is more than $1.5(13 - 6) = 10.5$ units beyond P_{75} .

—Kenneth O. McGraw

REFERENCE

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

INTERRATER AGREEMENT

Researchers often attempt to evaluate the consensus or agreement of judgments or decisions provided by a group of raters (or judges, observers, experts, diagnostic tools). The nature of the judgments can be nominal (e.g., Republicans/Democrats/Independents, or Yes/No); ordinal (e.g., low, medium, high); interval; or ratio (e.g., Target A is twice as heavy as Target B). Whatever the type of judgments, the common goal of interrater agreement indexes is to assess to what degree raters agree about the precise values of one or more attributes of a target; in other words, how much their ratings are interchangeable. Interrater agreement has been consistently confused with INTERRATER RELIABILITY both in practice and in research (Kozlowski & Hatrup, 1992; Tinsley & Weiss, 1975). These terms represent different concepts and require different measurement indexes. For instance, three reviewers, A, B, and C, rate a manuscript on four dimensions—clarity of writing, comprehensiveness of literature review, methodological adequacy, and contribution to the field—with six response categories ranging from 1 (*unacceptable*) to 6 (*outstanding*). The ratings of reviewers A, B, and C on the four dimensions are (1, 2, 3, 4), (2, 3, 4, 5), and (3, 4, 5, 6), respectively. The data clearly indicate that these reviewers completely disagree with one another (i.e., zero interrater agreement), although their ratings are perfectly consistent (i.e., perfect interrater reliability) across the dimensions.

There are many indexes of interrater agreement developed for different circumstances, such as different types of judgments (nominal, ordinal, etc.) as well as varying numbers of raters or attributes of the target being rated. Examples of these indexes are percentage agreement; STANDARD DEVIATION of the rating; STANDARD ERROR of the mean; Scott's (1955) π ; Cohen's (1960) kappa (κ) and Cohen's (1968) weighted kappa (κ_w); Lawlis and Lu's (1972) χ^2 test; Lawshe's (1975) Content Validity Ratio (CVR) and Content Validity Index (CVI); Tinsley and Weiss's (1975) T index; James, Demaree, and Wolf's (1984) $r_{WG(J)}$; and Lindell, Brandt, and Whitney's (1999) $r_{WG(J)}^*$. In the remaining section, we will review Scott's π , Cohen's κ and weighted κ_w , Tinsley and Weiss's T index, Lawshe's CVR and CVI, James et al.'s $r_{WG(J)}$, and Lindell et al.'s $r_{WG(J)}^*$. These indexes have been

widely used, and their corresponding NULL HYPOTHESIS tests have been well developed. Scott's π applies to categorical decisions about an attribute for a group of ratees based on two independent raters. It is defined as $\pi = \frac{p_o - p_e}{1 - p_e}$ and practically ranges from 0 to 1 (although negative values can be obtained), where p_o and p_e represent the percentage of observed agreement between two raters and the percentage of expected agreement by chance, respectively. Determined by the number of judgmental categories and the frequencies of categories used by both raters, p_e is defined as the sum of the squared proportions of overall categories,

$$p_e = \sum_{i=1}^k p_i^2,$$

where k is the total number of categories and p_i is the proportion of judgments that falls in the i th category.

Cohen's κ , the most widely used index, assesses agreement between two or more independent raters who make categorical decisions regarding an attribute. Similar to Scott's π , Cohen's κ is defined as $\kappa = \frac{p_o - p_e}{1 - p_e}$ and often ranges from 0 to 1. However, p_e in Cohen's κ is operationalized differently, instead as the sum of joint probabilities of the MARGINALS in the CONTINGENCY TABLE across all categories. In contrast to Cohen's κ , Cohen's κ_w takes disagreements into consideration. In some practical situations, such as personnel selection, disagreement between "definitely succeed" and "likely succeed" would be less critical than "definitely succeed" and "definitely fail." Both κ and κ_w tend to overcorrect for the chance of agreement when the number of raters increases.

Tinsley and Weiss's T index, $T = \frac{N_1 - N p_e}{N - N p_e}$, was developed based on Cohen's κ and Lawlis and Lu's χ^2 test of the RANDOMNESS in ratings, where N_1 , N , and p_e are the number of agreements across attributes, the total number of attributes rated, and the expected agreement by chance, respectively. Divided by N , the above formula can be rewritten as $T = \frac{p_o - p_e}{1 - p_e}$, where p_o is the proportion of attributes on which the raters agree. The value of p_e is determined by the range of acceptable ratings (r), the number of response categories used to make the judgments about the attributes (A), and the number of raters (K). If an exact agreement is required by researchers (i.e., the range of ratings is zero), $p_e = (\frac{1}{A})^{(K-1)}$. The computation formulas for p_e when $r > 0$ are quite complex; as such, a variety of p_e values can be obtained from Tinsley and Weiss (1975, note 7).

Lawshe's CVR, originally developed to assess experts' agreement about test item importance, is defined as $CVR = \frac{p_A - .5}{.5} = 2p_A - 1$ and may range from -1 to 1 , where p_A is the proportion of raters who agree. If raters judge multiple items, an overall interrater agreement index (CVI) is applied. CVI is defined as $CVI = 2\bar{p}_A - 1$, where $\bar{p}_A$ is the average of the items' p_A values.

James et al.'s $r_{WG(J)}$ compares the observed variance of a group of ratings to an expected variance from random responding. When one attribute is rated,

$$r_{WG(1)} = 1 - \frac{S_x^2}{S_E^2},$$

where S_x^2 is the observed variance for rating an attribute X and S_E^2 is the variance of random responding. Although there are numerous ways to determine S_E^2 , researchers often assume that the distribution of random responding is a discrete uniform distribution (Cohen, Doveh, & Eick, 2001). Given that, S_E^2 can be estimated by $S_{EU}^2 = \frac{A^2 - 1}{12}$, where A^2 is the number of response categories used to make judgments about attribute X . When there are J attributes to be rated,

$$r_{WG(J)} = \frac{J(1 - \frac{\bar{S}_x^2}{S_E^2})}{J(1 - \frac{\bar{S}_x^2}{S_E^2}) + \frac{\bar{S}_x^2}{S_E^2}},$$

where $\bar{S}_x^2$ is the average variance of J attributes. Similarly, S_E^2 can be estimated by S_{EU}^2 if the uniform distribution for random responding is assumed. Similar to the above indexes, the upper bound of $r_{WG(J)}$ is 1. If $r_{WG(J)}$ is negative, James et al. suggest setting it to zero.

Lindell et al.'s $r_{WG(J)}^*$ differs from James et al.'s $r_{WG(J)}$ on two grounds. First, they argued that interrater agreement that is greater than expected by chance is just as legitimate a concept as interrater agreement that is less than expected by chance (Lindell & Brandt, 1999, p. 642). By setting negative values of $r_{WG(J)}$ to zero, Cohen et al. (2001) found a positive bias. Second, James et al.'s $r_{WG(J)}$ is derived from the SPEARMAN-BROWN correction formula, although $r_{WG(J)}$ is an agreement index rather than a reliability index. Lindell et al. questioned the legitimacy of using the Spearman-Brown correction for an agreement index and recommended

$$r_{WG(J)}^* = 1 - \frac{\bar{S}_x^2}{S_E^2}$$

or

$$r_{WG(J)}^* = 1 - \frac{\bar{S}_x^2}{S_{EU}^2}$$

if the uniform distribution is assumed. In general, both $r_{WG(J)}$ and $r_{WG(J)}^*$ are more flexible in a variety of situations than Lawshe's CVR or CVI, as well as Tinsley and Weiss's T index.

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

- Cohen, J. (1960/1968). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the $r_{WG(J)}$ index of agreement. *Psychological Methods*, 6, 297–310.
- James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69, 85–98.
- Kozlowski, S. W., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology*, 77, 161–167.
- Lawlis, G. F., & Lu, E. (1972). Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin*, 78, 17–20.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563–575.
- Lindell, M. K., & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of the CVI, T , $r_{WG(J)}$, and $r_{WG(J)}^*$ indexes. *Journal of Applied Psychology*, 84, 640–647.
- Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement*, 23, 127–135.
- Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321–325.
- Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22, 358–376.

INTERRATER RELIABILITY

Behavioral scientists often need to evaluate the consistency of raters' judgments pertaining to characteristics of interest such as patients' aggressive behaviors in a geriatric unit, quality of submitted

manuscripts, artistic expression of figure skaters, or therapists' diagnostic abilities. Inconsistent judgments among raters threaten the validity of any inference extended from the judgment. Interrater reliability measures assess the relative consistency of judgments made by two or more raters.

Ratings made by two raters can be viewed as scores on two alternate forms of a test. Hence, interrater reliability based on two raters can be estimated by a SIMPLE CORRELATION such as PEARSON'S CORRELATION COEFFICIENT or SPEARMAN CORRELATION COEFFICIENT, according to the classical theory presented in RELIABILITY. However, these methods are of little use when there are more than two raters, such as in the case of a navy commander who needs to estimate the consistency of judgments among three officers who evaluate each F-18 pilot's landing performance. The above methods also cannot adequately assess interrater reliability if each pilot is rated by a different group of officers. To assess interrater reliability given the above circumstances, the most appropriate measure would be the INTRAClass CORRELATION (ICC), which is a special case of the one-facet GENERALIZABILITY study.

Conceptually, the ICC ranges from 0 to 1 in theory (although it is not always the case in reality) and refers to the ratio of the variance associated with ratees over the sum of the variance associated with ratees and error variance. Different ICC formulas are used for different circumstances. We will focus on three basic cases that are frequently encountered in both practice and research. For more complex scenarios, such as the assessment of interrater reliability for four raters across three different times, an application of generalizability theory should be utilized.

The three scenarios, well articulated in the seminal article by Shrout and Fleiss (1979), are as follows: (a) Each ratee in a random sample is independently rated by a *different* group of raters, which is randomly selected from a population; (b) each ratee in a random sample is independently rated by the *sample* group of raters, which is randomly selected from a population; and (c) each ratee in a random sample is independently rated by the sample group of raters, which contains the only raters of interest (i.e., there is little interest in whether the result can be generalized to the population of raters). Each case requires a different statistical analysis, described below, to calculate the correspondent ICC. Specifically, one-way random effects, two-way random effects, and two-way mixed models are appropriate for Case 1, Case 2, and Case 3,

Table 1 Rating Data Used to Demonstrate ICC for Case 1, Case 2, and Case 3

Ratee	Rater 1	Rater 2	Rater 3
A	2	4	7
B	2	4	7
C	3	5	8
D	3	5	8
E	4	6	9
F	4	6	9
G	5	7	10
H	5	7	10
I	6	8	11
J	6	8	11

Table 2 Analysis of Variance Result for Case 1

Sources of Variance	df	MS
Between Ratee	9	6.67
Within Ratee	20	6.33

respectively. Note that all of these ICC indexes are biased but consistent estimators of reliability.

In Case 1, we assume there are 10 ratees, and each ratee is rated by a different group of raters. Specifically, the three raters who rate Ratee C are different from those who rate Ratee E. Based on the data in Table 1 and the ANALYSIS OF VARIANCE result in Table 2, we can calculate the ICC by the formula

$$ICC(1, 1) = \frac{MS_b - MS_w}{[MS_b + (k - 1)MS_w]},$$

where MS_b and MS_w are the mean square between ratees and the mean square within ratees, respectively, and k is the number of raters rating each ratee. If there are different numbers of raters rating each ratee, k can be estimated by the average number of raters per ratee. The ICC (1,1), the expected reliability of a single rater's rating, is 0.02, which suggests that only 2% of the variation in the rating is explained by the ratee. If we are interested in the interrater reliability pertaining to the average rating of k raters, we can estimate it by

$$ICC(1, k) = \frac{MS_b - MS_w}{MS_b},$$

which is 0.05.

In the second case, there are 10 ratees who are rated by the same three raters. These raters are a random

Table 3 Analysis of Variance Result for Case 2 and Case 3

Sources of Variance	df	MS
Ratee	9	6.67
Rater	2	63.33
Residual	18	5.403E-15

sample of the rater population. According to Tables 1 and 3, we can calculate the ICC by the formula,

$$ICC(2, 1) = \frac{MS_b - MS_e}{[MS_b + (k - 1)MS_e + \frac{k(MS_j - MS_e)}{n}]},$$

where MS_b , MS_j , and MS_e are the mean square between ratees, the mean square between raters, and the mean square error respectively, k is the number of raters, and n is the number of ratees. Based on the sample data, the ICC(2, 1) is 0.35. If we are interested in the interrater reliability pertaining to the average rating of k raters, we can estimate it by

$$ICC(2, k) = \frac{MS_b - MS_e}{MS_b + \frac{MS_j - MS_e}{n}},$$

which is 0.51.

In the final case, there are 10 ratees who are rated by the same three raters. The difference between Case 2 and Case 3 is that we are interested in generalizing the result to the population of raters in Case 2, but we are interested only in the actual raters in Case 3. Similarly, we can use the information in Table 3 and calculate the ICC by the formula

$$ICC(3, 1) = \frac{MS_b - MS_e}{[MS_b + (k - 1)MS_e]},$$

and the ICC (3, 1) approximates 1. If we are interested in the interrater reliability pertaining to the average rating of k raters, we can estimate it by

$$ICC(3, k) = \frac{MS_b - MS_e}{MS_b}$$

only if there is no interaction between ratee and rater.

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.

INTERRUPTED TIME-SERIES DESIGN

The interrupted time-series design provides an estimate of the CAUSAL effect of a discrete intervention. In its simplest form, the design begins from a long series of repeated measurements on a DEPENDENT VARIABLE. The INTERVENTION breaks this TIME SERIES into preintervention and postintervention segments, and a data analysis compares the means of the dependent variable in the two periods. Rejecting the NULL HYPOTHESIS of no difference between the means is taken as evidence that the intervention influenced the series.

Figure 1 illustrates an unusually clear example of the design's logic. The data come from a 1978 study by A. John McSweeney and consist of 180 months of local directory assistance calls in Cincinnati, Ohio. During March 1974, the Cincinnati telephone company began to charge a small fee for answering these calls, which it had previously handled for free. Call volume plummeted, supporting the idea that the fee successfully reduced demands on the service.

In this example, the causal impact of the intervention seems obvious. The number of calls fell immediately and dramatically, and no ready explanation exists for why the drop might otherwise have occurred. More typically, any change in a series will be visually ambiguous, and alternative explanations will be easier to construct. Therefore, applications of the time-series design must address statistical analysis issues, and researchers must entertain rival explanations for an apparent causal effect.

Much of the design's popularity stems from the work of Donald T. Campbell, in collaborations with Julian C. Stanley and Thomas D. Cook. Campbell and his coworkers developed a list of threats to the INTERNAL VALIDITY of causal inferences, where variables besides the one under study might account for the results. Using their list to evaluate common nonexperimental designs, Campbell and coworkers concluded that the interrupted time series had stronger internal validity than did most alternatives.

Its relative strengths notwithstanding, the time-series design is still vulnerable to several important validity threats. One of these is *history*, or the possibility that other events around the time of the intervention were responsible for an observed effect. To make historical threats less plausible, researchers often analyze control series that were not subject to the intervention.

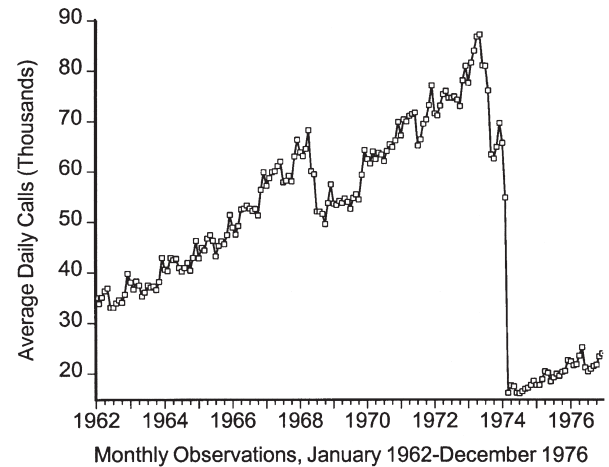


Figure 1 Average Daily Calls (in Thousands) to Cincinnati Local Directory Assistance, January 1962–December 1976

SOURCE: McSweeney (1978).

Lack of change in these series increases confidence that the intervention was the causal agent. In Cincinnati, for example, the telephone company continued to provide free long-distance directory assistance, and this series did not drop after the local assistance fee began.

Statistical analyses of interrupted time-series data often rely on BOX-JENKINS MODELING. Although not absolutely required, Box-Jenkins methods help solve two major analytic problems: nonstationarity and SERIAL CORRELATION. Nonstationarity occurs when a series lacks a constant mean (e.g., when it trends), and serial correlation occurs when later observations are predictable from earlier ones. If not addressed, these features can invalidate tests of the difference between the preintervention and postintervention means.

—David McDowall

REFERENCES

- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.
- McDowall, D., McCleary, R., Meidinger, E. E., & Hay, R. A., Jr. (1980). *Interrupted time series analysis*. Beverly Hills, CA: Sage.
- McSweeney, A. J. (1978). Effects of response cost on the behavior of a million persons: Charging for directory assistance in Cincinnati. *Journal of Applied Behavior Analysis*, 11, 47–51.

INTERVAL

An interval is a level of measurement whereby the numeric values of the variable are equidistant and have intrinsic meaning. A classic example is income measured in dollars. An income score of \$90,000 is immediately understood. Furthermore, the distance between someone with a \$90,000 income and someone with a \$90,001 income (i.e., one dollar) is equal to the distance between someone with a \$56,545 income and someone with a \$56,546 income (i.e., one dollar). Interval level of MEASUREMENT allows a high degree of precision and the ready use of MULTIVARIATE statistics. Interval measures are sometimes called QUANTITATIVE, or METRIC, measures. DATA that are interval can be converted to lower levels of measurement, such as ORDINAL or NOMINAL, but information is lost.

—Michael S. Lewis-Beck

INTERVENING VARIABLE. See
MEDIATING VARIABLE

INTERVENTION ANALYSIS

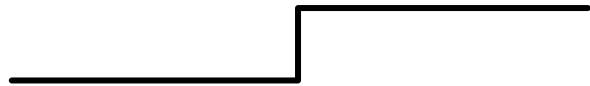
Interventions in a time series are measures representing disruptions in the course of a series of data extending over a time period. They are often conceptualized for evaluating policy impacts or other discrete introductions in a historical process. In many cases, they are introduced as the experimental treatment in an INTERRUPTED TIME-SERIES DESIGN (McDowall, McCleary, Meidinger, & Hay, 1980). Because estimates of policy interventions or other factors affecting a time series can be BIASED by contaminating factors, such as TREND (nonstationarity), AUTOREGRESSION (sometimes present even when there is no trend, as in cyclical processes), and MOVING AVERAGE components (random shocks that persist over time and then die away), the analytical model is usually specified as the presence or absence of an intervention (0 or 1), plus a noise model from which unspecified time-related effects have been eliminated—a WHITE NOISE process. The equation for this process is usually represented as

$$y_t = \omega I_t + N_t,$$

1. Pulse Function ($I = \dots, 0, 0, 1, 0, 0, \dots$)



2. Step Function ($I = \dots, 0, 0, 1, 1, 1, \dots$)



3. Ramp Function ($I = \dots, 0, 0, 1, 2, 3, 4, \dots$)

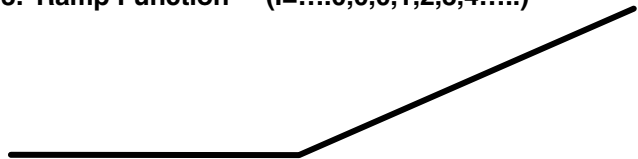
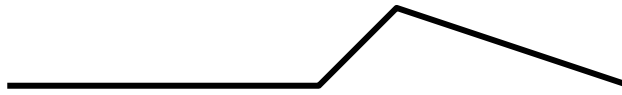


Figure 1 Basic Structures of Interventions

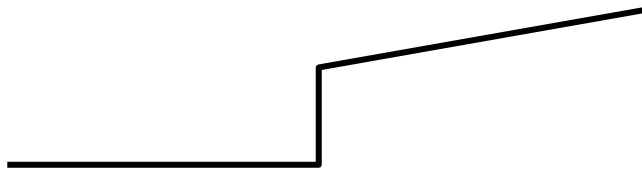
where y_t represents an observation in a time series composed of ωI_t ; effects of the intervention; and a *white noise process*, N_t . Effects of the intervention can be evaluated by a T -TEST for SIGNIFICANCE of ω . This has the effect of eliminating time-series components extraneous to the impact of the intervention and permits testing of hypotheses regarding impacts of the policy or other intervention.

Although the research takes the form of a QUASI-EXPERIMENTAL design suggested by Campbell and Stanley (see INTERRUPTED TIME-SERIES DESIGN), estimation of the intervention effects may take a variety of specifications. When precise information on the form of the intervention is available, such as yearly reductions in allowable pollution (Box & Tiao, 1975), the intervention should be specified accordingly. Most policy interventions are difficult to specify, however, and require DUMMY VARIABLES representing presence or absence of a policy or program.

1. Rapidly Dampening Pulse Function ($I = \dots 0, 0, 0, 1, 0, 0, \dots$; $\delta > 0$)



2. Step Function Plus Gradually Increasing Level ($I = \dots 0, 0, 0, 1, 1, 1, \dots$; $\delta > 0$)



Model 2 can also be modeled as a Step Function plus a Ramp Function by specifying two intervention effects.

3. Step Function Plus Ramp Function ($I_1 = \dots 0, 0, 0, 1, 1, 1, \dots$; $I_2 = \dots 0, 0, 0, 1, 2, 3, 4, \dots$)

Figure 2 Common, Complex Models of Intervention Effects

For these more general effects, there are three basic specification models that account for an intervention (Figure 1).

Tests of these models can be obtained by including a lag of y , y_{t-1} , in the estimating equation,

$$y_t = \delta y_{t-1} + \omega I_t + N_t,$$

and evaluating the parameter, δ , by a t -test. If the analyst has no explicit knowledge as to the form of the intervention, alternative specifications can be tested. Specifically, one may try specification of the intervention in the following manner:

1. If the parameter of a specified *pulse function* (abrupt, temporary effect) is significant and the parameter, δ , is not significant, the intervention effect takes the form of the *pulse function* illustrated in Figure 1.
2. If the parameter of a specified *pulse function* (abrupt, temporary effect), ω , is significant and the parameter, δ , is also significant, the outcome suggests that a *step function* (abrupt, permanent effect) is a better specification.

For example, if the coefficient of the lagged endogenous variable equals 1, the model is identical with a *step function*.

3. If the parameter of a specified *step function*, ω , is significant, but the parameter, δ , is not significant, a *step function* specification is probably correct (Figure 1).
4. If the parameters of both the intervention specified as a *step function* and the lagged time series, y_{t-1} , prove to be significant, alternative model combinations should be investigated.
5. If the parameter of a *pulse function* specification is not significant, but the parameter of y_{t-1} is significant, a *ramp function* (Figure 1) should be investigated.

When ambiguities are detected in the model specifications described above, the model may require more complex, alternative specifications, sometimes as combinations of the basic models (see Figure 2).

These models of interventions represent the most frequently encountered specifications. More complex models, including multiple interventions, are also useful for representing policy and program effects.

Scholars sometimes object that following the Box and Jenkins specification of a white noise process by removing time-related processes is atheoretical. If a researcher believes that specifiable variables may represent alternative rival hypotheses, they may be included as transfer functions. In any case, the process is hardly more atheoretical than so-called ERROR CORRECTION models and has the effect of eliminating contaminating statistical artifacts, such as COLLINEARITY, from the equation.

—Robert B. Albritton

See also INTERRUPTED TIME-SERIES DESIGN

REFERENCES

- Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.
- Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70, 70–92.
- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.
- McDowall, D., McCleary, R., Meidinger, E. E., & Hay, R. A., Jr. (1980). *Interrupted time series analysis*. Beverly Hills, CA: Sage.

INTERVIEW GUIDE

An interview guide, or aide memoire, is a list of topics, themes, or areas to be covered in a SEMISTRUCTURED INTERVIEW. This is normally created in advance of the interview by the researcher and is constructed in such a way as to allow flexibility and fluidity in the topics and areas that are to be covered, the way they are to be approached with each interviewee, and their sequence. The interview guide normally will be linked to the RESEARCH QUESTIONS that guide the study and will cover areas likely to generate data that can address those questions.

An interview guide is a mechanism to help the interviewer conduct an effective semistructured interview. It can take a variety of forms, but the primary consideration is that it enables the interviewer to ask questions relevant to the research focus, in ways relevant to the interviewee, and at appropriate points in the developing social interaction of the interview. An interview guide should help an interviewer to make

on-the-spot decisions about the content and sequence of the interview as it happens. This means, for example, deciding whether a topic introduced by the interviewee is worth following up, when to close a topic and open a new one, whether and when to reintroduce something mentioned earlier in the interview, and when and how to probe for more detail. It also usually means deciding on the spot how to word a question in the specific context of a particular interview, rather than reading from a script.

This level of flexibility and specificity cannot be achieved with a standardized list of interview questions, and thus an interview guide should be distinguished from an INTERVIEW SCHEDULE, which contains a formal list of the questions to be asked of each interviewee. Instead, an interview guide might comprise topics headings and key words on cards that can easily be shuffled and reordered. It might include examples of ways of asking about key issues, but it will not list every or even any question in detail. It might include some of the researcher's hunches, or specific issues or events that would be worth probing should they arise. It might include a checklist of areas, or perhaps interviewee characteristics, to ask about at the end of the interview if these have not arisen before.

The quality of an interview guide is dependent upon how effective it is for the researcher and the research in question, rather than upon abstract principles. Therefore, it is vital that interview guides are tested out in a PILOT STUDY and modified accordingly.

—Jennifer Mason

REFERENCES

- Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage.
- Mason, J. (2002). *Qualitative researching* (2nd ed.). London: Sage.

INTERVIEW SCHEDULE

An interview schedule is the guide an interviewer uses when conducting a STRUCTURED INTERVIEW. It has two components: a set of questions designed to be asked exactly as worded, and instructions to the interviewer about how to proceed through the questions.

The questions appear in the order in which they are to be asked. The questions are designed so they can be administered verbatim, exactly as they are written. The questions need to communicate not only what information is being asked of RESPONDENTS but also the form or the way in which respondents are supposed to answer the question. For example, "Age" does not constitute a question for an interviewer to ask. In a structured interview schedule, the question would supply all the words the interviewer would need: "How old were you on your last birthday?"

Similarly, "When were you born?" does not tell the respondent what kind of answer to provide. "After the Vietnam war," "about 20 years ago," and "in 1971" would all constitute adequate answers to this question, yet they are all different. A good question in an interview schedule would read, "In what year were you born?"

The interview schedule also needs to make clear to interviewers how they are to proceed through the questions. For example:

1. Provide optional wording to fit particular circumstances of the respondent. For example, "Did you (or your husband/wife) buy any stocks in the past 12 months?" The parentheses and the husband/wife wording communicates to interviewers that they have to adapt the question to the respondent's situation. This helps the interviewer word questions properly and ensures that all interviewers adapt question wording in the same, consistent way.
2. Provide instructions for skipping questions. For example, certain questions might be asked only of those who have recently visited a doctor. The interview schedule must identify those people who have and have not been to a doctor's office and provide appropriate instructions to interviewers about which questions to ask and which to skip.
3. Interview schedules also provide instructions to interviewers about where and how to record answers.

Interview schedules can be on paper forms in booklets. Increasingly, interview schedules are computerized for COMPUTER-ASSISTED DATA COLLECTION; the questions to be asked pop up on the computer screen to be read by the interviewer, and answers are recorded by either typing the respondent's words into the computer or by entering a number that corresponds to a particular answer. In either form, the role of the interview

schedule is to provide a PROTOCOL for interviewers to ask and record answers in a consistent way across all respondents and to facilitate the process of getting through the interview smoothly and efficiently.

—Floyd J. Fowler, Jr.

REFERENCES

- Converse, J., & Presser, S. (1986). *Survey questions*. Beverly Hills, CA: Sage.
- Fowler, F. J., Jr. (1995). *Improving survey questions*. Thousand Oaks, CA: Sage.
- Sudman, S., & Bradburn, N. (1982). *Asking questions*. San Francisco: Jossey-Bass.

INTERVIEWER EFFECTS

The increase in VARIANCE of a SAMPLE statistic because of interviewer differences is known as the *interviewer effect*. This increased variance can be modeled as arising from a positive CORRELATION, ρ_{int} , among responses collected by the same interviewer. Under this simple MODEL, it can be shown that variance of the sample mean is increased by a factor of about $1 + (\bar{m} - 1)\rho_{\text{int}}$, where $\bar{m}$ is the size of the average interviewer assignment (Kish, 1962).

Studies have shown that ρ_{int} is typically small, often less than 0.02 and nearly always less than 0.10. Even so, it can have a large impact on the STANDARD ERROR of statistics when interviewer assignments are large. For example, the standard error of the sample MEAN of an item having $\rho_{\text{int}} = 0.02$ would be nearly doubled in a survey in which the average interviewer assignment size is 150 because its variance would be increased by a factor of $1 + (150 - 1)(0.02) = 3.98$. This suggests that one way to reduce the impact of interviewer differences on survey data is to increase the number of interviewers on staff and reduce their average assignment size.

Interviewer effects are not routinely measured as part of survey operations. To do so, one must interpenetrate interviewer assignments, meaning that the sample units must be RANDOMLY ASSIGNED to them. This is prohibitively expensive in personal visit surveys, but more feasible in telephone survey operations. Even there, though, potential shift effects complicate ESTIMATION. A methodology that allows interviewer and respondent covariates to be accounted for, called MULTILEVEL ANALYSIS, has been adopted for modeling

interviewer effects. Software to implement this method is now readily available (e.g., MLwiN) and could help make estimation of interviewer effects in production environments more feasible.

One advantage of measuring interviewer effects is that it allows a more realistic assessment of precision of estimates of means and totals. Another is that it provides a way of assessing and comparing data quality among the survey items. Ideally, this information would help with monitoring and improving interviewer performance. However, it is still not clear how to reduce the magnitude of ρ_{int} for an item, even if it is known to be large. Numerous studies have attempted to find the causes of interviewer effects. Demographic and behavioral characteristics of interviewers, such as race, gender, voice, and demeanor, have been examined. Little evidence has been found to suggest that these characteristics explain the effects. Other research has tried to identify the types of items most affected. Although it seems plausible that sensitive items would be more vulnerable, this has not been clearly shown. There is some evidence that items requiring more interviewer intervention, such as PROBING, have larger interviewer effects, which suggests that more extensive interviewer training in these skills might be useful. However, some studies have suggested that training alone may not be helpful, but the monitoring and feedback possible in a centralized telephone facility may help. Until recently, it was widely accepted that standardized interviewing, which restricts interviewers to a uniform script, was the best approach for controlling interviewer effects. Some research now suggests that a more conversational style, which allows interviewers to paraphrase or clarify responses, may increase the accuracy of responses without increasing ρ_{int} (Schober & Conrad, 1997).

—S. Lynne Stokes

REFERENCES

- Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (Eds.). (1991). *Measurement errors in surveys*. New York: Wiley.
- Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92–115.
- Schober, M., & Conrad, F. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576–602.

INTERVIEWER TRAINING

Those who are going to be interviewers on social science research projects almost always benefit from training. The specific mix of skills and techniques they need to learn varies somewhat by whether they are doing STRUCTURED or SEMISTRUCTURED INTERVIEWS.

If interviewers will be doing structured interviews, the basic skills they need to learn include the following:

1. Asking questions exactly as worded
2. Probing incomplete answers in a nondirective way; that is, in a way that does not increase the likelihood of one particular answer over others
3. Recording answers without discretion

When interviewers are doing SEMISTRUCTURED INTERVIEWS, they need to be trained in how to word or ask questions in a way that will elicit the information they need to obtain. The answers to semistructured interviews tend to be in narrative form in the respondents' own words, whereas structured interviews tend to ask respondents to choose one of a specific set of answers. Therefore, semistructured interviewers also have to have more highly developed skills in nondirective probing than is typically the case for a structured interviewer. Specifically, such interviewers need to develop skills at diagnosing the reason why an answer is incomplete and learn an array of nondirective probes that will elicit the information that is needed.

Two skills are common to both kinds of interviewers. First, interviewers usually are responsible for enlisting the cooperation of potential respondents. Thus, they need to learn how to introduce the research project in a way that engages the respondents' interest and makes them willing to accept their role in the interview. Second, interviewers also have to train the respondents in what they are supposed to do during the interview. This is particularly a challenge for structured interviews, in which the behavior of both interviewers and respondents is highly prescribed. For such interviews, it is extremely valuable for interviewers to properly explain to respondents what they are supposed to do and why it is the best way to do it, in order to get them to accept their role and play it well. In contrast, a semistructured interview is less constrained, so training the respondents in their roles is less of a challenge.

There have been several studies that show that the training of structured interviewers has a major effect on how well they conduct interviews and on the quality of data that they collect. Interviewers with minimal training do not perform well. In particular, it has been shown that supervised practice interviewing has a positive effect on interviewing and data quality. For semistructured interviews, we lack systematic studies of the effects of training on the resulting data. However, studies have shown that PROBING answers that are provided in narrative form is among the hardest skills for interviewers to master, and those skills benefit most from training. The same studies suggest that designing appropriate question wording may be even harder and, hence, even more likely to benefit from training.

—Floyd J. Fowler, Jr.

REFERENCES

- Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized survey interviewing*. Newbury Park, CA: Sage.
- Weiss, R. (1995). *Learning from strangers*. New York: Free Press.

INTERVIEWING

In the social sciences, a major method of data gathering is interviewing individuals. Interviews may be formally conducted in social SURVEYS, face-to-face or over the telephone. Or, subjects in a study population, such as members of an indigenous language community in the Andes, can be more or less informally interviewed. At the elite level, interviews may be carried out essentially as conversations, at least ostensibly. The interview method may have to be adapted to the social context in order to achieve the overriding goal of obtaining VALID and useful DATA.

—Michael S. Lewis-Beck

INTERVIEWING IN QUALITATIVE RESEARCH

INTERVIEWING is one of the most frequently used research methods in the social sciences, used in both quantitative SURVEY research and in QUALITATIVE RESEARCH. Qualitative interviewing involves a special

kind of conversation, one in which an interviewer (or more than one) asks questions of a respondent or subject (or more than one), on a particular topic or topics, and carefully listens to and records the answers. Some interviews take place over the telephone, and some in person; some are videotaped, whereas most are audiotaped and then transcribed. This discussion of the qualitative interview is based on what is probably the most common form: a single interviewer interviewing a single respondent using an audiotape recorder.

Interviews have been used in the social sciences since the 19th century, when scholars and practitioners began to seek answers to large-scale questions about the human condition. In America, interviews were used as part of the case study method in the social sciences at the University of Chicago during the 1920s and 1930s, together with fieldwork, documentary analysis, and statistical methods (Platt, 2002). After World War II, interviews began to be used extensively in survey research; from that point onward, the quantifiable CLOSED-ENDED QUESTION survey interview became more common than the open-ended qualitative interview.

However, qualitative interviewing remained as a key method in anthropology and is used today in all the social sciences. The purpose of qualitative interviewing in social science research today, as of qualitative research in general, is to understand the meanings of the topic of the interview to the respondent. As Kvale (1996) puts it, the qualitative interview has as its purpose “to obtain descriptions of the life world of the interviewee with respect to interpreting the meaning of the described phenomena” (pp. 5–6). For the sake of simplicity, I illustrate this point mainly with reference to my own qualitative interview-based book on Madwives (Warren, 1987).

In interviews with female psychiatric patients in the late 1950s and early 1960s (Warren, 1987), the eight interviewers wanted to know about the experiences of mental hospitalization from the perspective of the female patients as well as the psychiatrists. They interviewed these women at weekly intervals during and after hospitalization, in some cases for several years. One of the topics they asked about was electroshock or electroconvulsive therapy (ECT). From the point of view of the medical model, ECT is a psychiatric treatment intended to relieve the symptoms of depression, schizophrenia, or other disorders, with side effects that include short-term memory loss. To the women undergoing ECT in the late 1950s and early 1960s, however,

the meanings of the treatment were both different (from the medical model) and variable (from one another, and sometimes from one day to another). Some of the women saw it as a form of punishment, whereas others believed that it was a medical treatment, but that its purpose was to erase their troubles from their memories.

Social scientists interview people because they have questions about the life worlds of their respondents. How do mental patients interpret the treatments they are given? What do single women involved in affairs with married men think about their situation (Richardson, 1985)? In order to use the qualitative interview method, these RESEARCH QUESTIONS have to be translated into actual questions for the researcher to ask the respondent. These questions are developed in two contexts: the overall RESEARCH DESIGN, and governmental and institutional guidelines for protection of human subjects.

In social research, qualitative interviewing may be part of a research design that also includes FIELD RESEARCH or other methods, or it may be the only method planned by the researcher. Whatever the overall design, the specific questions asked by the interviewer are related to the topic of interest and to the study designer's research questions. In general, qualitative researchers develop 10 to 15 questions to ask respondents, and they are also prepared to use PROBES to elicit further information. In order to save time, basic demographic information such as gender, age, occupation, and ethnicity can be gathered using a face sheet for each respondent. Then, the rest of the interview time—which may range from 15 minutes to several hours, depending upon the questioner and the talkativeness of the respondent—can be spent on the researcher's topic questions and the respondent's narrative.

In contemporary social science research, qualitative interview studies must concern themselves with federal and institutional human subjects regulations. Although human subjects regulations are discussed elsewhere in this volume, it is important to note the requirements for INFORMED CONSENT and CONFIDENTIALITY in interview research. Generally, the respondents in an interview study are given a letter explaining the purpose of the research (informed consent) and indicating who is sponsoring it—a university department, for example—which they will sign. As with any other social research, the interviewer must promise (and keep that promise!) to make sure that the names and identifying details of all respondents remain confidential.

Once the interviewer has his or her Institutional Review Board (IRB) permissions, has his or her questions formulated, and is ready to go, he or she will have to find people to be interviewed—deciding, first, what types of people to interview (by gender, race, age, and so on), and how many (generally, 20–50). Then, the interviewer must locate specific respondents. The general issue of sampling is discussed elsewhere in this volume; suffice it to say here that the difficulty of finding people to talk to is dependent, in part, on the topic of the interview. If the study is to be on mental patients, for example, the researcher must locate a mental hospital that will give him or her permission to do the research (very difficult in the 2000s), and then gain individual patients' consent—both of which are fraught with difficulty. The study of single women having affairs with married men will probably involve fewer problems, especially if the researcher is a woman around the same age as the respondents she wants to interview. As Richardson (1985) notes, "Finding Other Women to interview was not difficult" (p. x), and getting them to tell their stories was not either, because women in that situation are often eager to talk.

Qualitative interviewing is a special kind of conversation, with the researcher asking the questions and the respondent taking as much time as he or she needs (within some limits!) to tell his or her story; qualitative interviews typically last from a half hour to an hour or more. The types of questions asked include introducing questions ("Can you tell me about . . .?"), probing questions ("Could you tell me more about . . .?"), and specifying questions ("What did you think then . . .?") (Kvale, 1996, pp. 133–134). The interviewer tries not to lead the witness; for example, instead of saying, "Didn't that make you feel awful?" he or she would ask, "How did that make you feel?" The order of questioning proceeds from the most general and unthreatening to, at the end, questions that might cause discomfort to the respondent. The interviewer hopes to gain some rapport with the respondent in the early part of the interview so that more sensitive questions will be acceptable later on.

Audiotaped qualitative interviews are generally TRANSCRIBED, either by the interviewer or by someone else. It is often preferable for the interviewer to transcribe the interviews he or she has done, because he or she is in the best position to be able to interpret ambiguous statements. Paid—or especially unpaid—transcribers lack the firsthand experience with the

interview and may lack the incentive to translate the oral into the written word with attention and care. It is important in transcribing and analyzing interviews to remember that an interview is a speech event guided by conversational turn-taking (Mishler, 1986).

Transcripts are used in qualitative interviewing—just as field notes are used in ETHNOGRAPHY—as the textual basis for analysis. Analysis of interview or ethnographic text consists of reading and rereading transcriptions and identifying conceptual patterns. In my study of female mental patients and their husbands in the late 1950s and early 1960s, I noticed two patterns in those interviews that occurred after the women had been released from the mental hospital. Some wives, and some husbands, wanted to forget and deny that the women had been in the mental hospital; the continued weekly presence of the interviewer, however, reminded them of this. So, they either tried to end the interviews, or, when that failed (respondents are sometimes very obedient!), they tried to make the interviews into social events by inviting the interviewer to stay for tea, or by turning the tables and asking questions. Other women were still very troubled and wanted to continue to try to make the interview process into a form of therapy (Warren, 1987). Here are some examples:

Ann Rand repeated that she would only see the interviewer again if she would have her over to her house. While the interviewer was evasive, Ann said, “then I suppose you still see me as a patient. To me you are either a friend or some kind of authority, now which is it? The way I see it, you either see me as a friend or a patient.” (p. 261)

Mr. Oren asked me if I wanted to join them for dinner, and was rather insistent about this despite my repeated declining. He later invited me to drop over to his place some afternoon, bring along a broad and just let my hair down and enjoy myself. . . . I was emphasizing the fact that this was my job, probably in the hope of getting him to accept the situation as such and not redefine it. (p. 261)

It is no accident that the first example involves a female interviewer and a female respondent, and the second involves a male interviewer and a male respondent. A housewife in 1960 would not have tried to make a friend out of a male interviewer—indeed, the male

interviewers had difficulty setting up interviews with female respondents—and a husband from the same era would not have invited a female interviewer to “drop over to his place and bring along a broad.” The extensive literature on qualitative interviewing documents the significance of gender and other personal characteristics on the interview process—and thus the data that are the subject of analysis. Some feminist qualitative interviewers in the 1970s and 1980s, such as Ann Oakley, noted the particular suitability of interviewing as a method for eliciting the narratives of women (whose everyday lives were often invisible) and for lessening the hierarchical divisions between superordinate researchers and subordinated subjects in social science research (Warren, 2002). Other feminist interviewers in the 1990s and 2000s, such as Donna Luff, note that there are profound differences between women in terms of nationality, race, social class, and political ideology—differences that are not necessarily overcome by the commonalities of gender (Warren, 2002).

Issues of representation and POSTMODERNISM are as relevant to qualitative interviewing in the social sciences as they are to ethnography and other methods. Gubrium and Holstein (2002) have challenged the modernist (and journalistic) notion of interviewees as “vessels of answers” whose responses are just waiting to be drained out of them by interviewers. Gubrium and Holstein’s and others’ critiques of the interview method (whether these critiques are from postmodernist, feminist, or other stances) focus on the EPISTEMOLOGY of the interview: What kind of knowing is it, and what kinds of knowledge can be derived from it? Various answers to these questions have emerged during the past decade or so, including the framing of interviews as a particular form of talk (Mishler, 1986) with knowledge of linguistic forms to be derived from it—and nothing else. Whether an interview is just a speech event, or can be used as data for understanding respondents’ meaning-worlds outside the interview, is an open question at the present time.

Although it is one of the most widely used methods in the social sciences, the qualitative interview is currently subject to considerable epistemological debate. As Gubrium and Holstein (2002) note, we live in an “interview society” in which everyone’s point of view is seen as worthy of eliciting, and in which everyone seemingly wants to be interviewed—at least by the media, if not by social scientists. The qualitative interview, like the survey interview in quantitative research,

is likely to be used for some time to come to illuminate those aspects of social lives that cannot be understood either quantitatively or ethnographically.

What is it that the qualitative interview can capture that cannot be captured by field research? Although ethnography does, like qualitative interviewing, seek an understanding of members' life-worlds, it does so somewhat differently. Going back to the example of mental hospital research, a field researcher observes staff-patient and patient-patient interaction in the present, coming to conclusions about the patterns and meanings observed in those interactions. An interviewer elicits biographical meanings related to the present but also the past and future; the interview focuses not on interaction (which is best studied observationally) but on accounts or stories about the respondent. Thus, to a field researcher, ECT looks like an easy and inexpensive way for hospitals to subdue unruly patients, whereas to the housewife undergoing ECT, it seems like a punishment for her inability to be what the culture tells her is a good wife and mother.

Interviewing in today's society holds both dangers and promises for the conduct of social science research. Interviewing is an inexpensive and easy way to do research in an interview society where quick results are the bottom line—especially for journalists—producing superficial and cursory results. But it can also be an invaluable tool for understanding the ways in which people live in and construct their everyday lives and social worlds. Without such an understanding, it is impossible to conduct good or useful theoretical, quantitative, or applied research. Whatever the aims of the social scientist—constructing or proving theory, conducting large-scale survey research, or improving social conditions—none of them can be reached without an understanding of the lives of the people who live in the society and the world. Qualitative interviewing, together with ethnography, provides the tools for such an understanding.

—Carol A. B. Warren

REFERENCES

- Gubrium, J. F., & Holstein, J. A. (2002). From the individual interview to the interview society. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research: Context and method* (pp. 3–32). Thousand Oaks, CA: Sage.
- Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage.
- Mishler, E. G. (1986). *Research interviewing: Context and narrative*. Cambridge, MA: Harvard University Press.
- Platt, J. (2002). The history of the interview. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research: Context and method* (pp. 33–54). Thousand Oaks, CA: Sage.
- Richardson, L. (1985). *The new other woman: Contemporary single women in affairs with married men*. London: Collier-Macmillan.
- Warren, C. A. B. (1987). *Madwives: Schizophrenic women in the 1950s*. New Brunswick, NJ: Rutgers University Press.
- Warren, C. A. B. (2002). Qualitative interviewing. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research: Context and method* (pp. 83–101). Thousand Oaks, CA: Sage.

INTRACLAS CORRELATION

An intraclass correlation coefficient is a measure of association based on a variance ratio. Formally, it is the ratio of the variance for an object of measurement (σ_{Object}^2) to the object variance plus error variance:

$$(\sigma_{\text{error}}^2) \frac{\sigma_{\text{Object}}^2}{\sigma_{\text{Object}}^2 + \sigma_{\text{error}}^2}.$$

It differs from an analog measure of association known as OMEGA SQUARED (ω^2) only in the statistical model used to develop the sample estimate of σ_{Object}^2 . Intraclass correlation estimates are made using the random effects model (i.e., the objects in any one sample are randomly sampled from a population of potential objects) and ω^2 estimates are made using the FIXED EFFECTS MODEL (i.e., the objects in the sample are the only ones of interest and could not be replaced without altering the question being addressed in the research).

Intraclass correlation is best introduced using one of the problems it was developed to solve—what the correlation of siblings is on a biometric trait. If the sample used to obtain the correlation estimate consisted just of sibling pairs (in which sibships are the object of measurement), a Pearson product-moment, interclass correlation could be used, but this procedure would not lead to a unique estimate. The reason is seen with an example. For three sibling pairs {1, 2}, {3, 4}, and {5, 6}, the Pearson r for pairs when listed in this original order {1, 2}{3, 4}{5, 6} is different from the Pearson r for the same pairs if listed as {2, 1}{3, 4}{6, 5}, even though the sibling data are exactly the same. The Pearson procedure requires the totally arbitrary assignment of pair members to a set

X and a set Y when, in fact, the only assignment is to sibship. How, then, is one to obtain a measure of correlation within sibship classes without affecting the estimate through arbitrary assignments? An early answer was to double-list each pair and then compute a Pearson r . For the example above, this meant computing the Pearson r for the twice-listed pairs {1, 2}{2, 1}{3, 4}{4, 3}{5, 6}{6, 5}.

Double-listing the data for each sibship leads to a convenient solution in the case of sibling pairs, but the method becomes increasingly laborious as sibships are expanded to include 3, 4, or k siblings. A computational advance was introduced by J. A. Harris (1913), and then R. A. Fisher (1925) applied his analysis of variance to the problem. This latter approach revealed that a sibling correlation is an application of the one-way random effects variance model. In this model, each sibling value is conceptualized as the sum of three independent components: a population mean (μ) for the biometric trait in question; an effect for the sibship (α), which is a deviation of the sibship mean from the population mean; and a deviation of the individual sibling from the sibship mean (ε). In its conventional form, the model is written $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$. The sibling correlation is then the variance of the alpha effects over the variance of the alpha effects plus epsilon (error) effects:

$$\frac{\sigma_{\alpha}^2}{\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2}.$$

In other words, it is the proportion of total variance in the trait that is due to differences between sibships. For traits that run in families, the correlation would be high; for those that don't, the correlation would be low.

Applying the ANALYSIS OF VARIANCE to the problem of sibling correlation solved the computational inconveniences of the old method, easily accommodated the analysis of data with unequal sibling sizes, and clarified the nature of the correlation as one of the ratio of variances. It also provided a means of computing other forms of intraclass correlation. Sibling correlations represent correlations expressible using a one-way model. There are other correlations of interest that require two- and three-way models. Many of these involve the RELIABILITY of measurements: How reliable are judges' ratings when each of k judges rates each of n performers? How reliably do school students perform across different instructors and subjects? Each of these questions can be addressed by forming variance ratios that are intraclass correlations. In fact, two of the most common reliability

measures—CRONBACH'S ALPHA and Cohen's kappa—are expressible as intraclass correlations.

—Kenneth O. McGraw

REFERENCES

- Fisher, R. A. (1925). *Statistical methods for research workers* (1st ed.). Edinburgh, UK: Oliver & Boyd.
- Harris, J. A. (1913). On the calculation of intraclass and interclass coefficients of correlation from class moments when the number of possible combinations is large. *Biometrika*, 9, 446–472.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.

INTRACODER RELIABILITY

An observer's or coder's judgments about people's behaviors or other phenomena of interest (e.g., violent acts on TV shows or gender stereotypes embedded in children's books) tend to fluctuate across different occasions. Intracoder (intrarater or intraobserver) reliability provides an estimate of the relative consistency of judgments within a coder over time. More specifically, it assesses the amount of inconsistency (i.e., MEASUREMENT ERRORS) resulting from factors such as carefulness, mood, noise, fatigue, and fluctuation of targets' behaviors that occurs during a period of time. Intracoder reliability differs from INTERRATER RELIABILITY in that the latter assesses the relative consistency of judgments made by *two or more* raters (or coders).

The conventional procedure to estimate intracoder reliability is similar to that of TEST-RETEST RELIABILITY, as described below. A coder makes a judgment about targets' behaviors (e.g., number of violent scenes on TV) at two different occasions that are separated by a certain amount of time. The two sets of judgments are then correlated by using various formulas discussed in SYMMETRIC MEASURES. This correlation coefficient represents an estimate of intracoder reliability.

In the remaining section, we will introduce a flexible but less used approach to assess intracoder reliability across multiple occasions based on GENERALIZABILITY THEORY, or G theory (Brennan, 1983; Cronbach,

Table 1 Number of Eye Contacts of 10 Autistic Pupils Observed by One Coder

	Occasions				Pupils	
	Day 1	Day 2	Day 3	Day 4	Mean	SD
Pupil A	0	2	2	0	1	1.15
Pupil B	2	3	1	1	1.75	0.96
Pupil C	2	0	1	2	1.25	0.96
Pupil D	1	3	0	0	1	1.41
Pupil E	1	0	0	1	0.5	0.58
Pupil F	4	3	2	2	2.75	0.96
Pupil G	1	0	3	0	1	1.41
Pupil H	2	1	0	3	1.5	1.29
Pupil I	3	1	2	0	1.5	1.29
Pupil J	3	1	0	0	1	1.41
Mean	1.9	1.4	1.1	0.9	Overall Mean: 1.33	
SD	1.2	1.26	1.1	1.1	Overall SD: 1.19	

Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). Let's assume a coder has observed the number of eye contacts of 10 autistic pupils 1 hour a day for 4 consecutive days. The observation records are presented in Table 1.

Based on the content in Table 1, we can identify four sources that are responsible for the variation in the coder's observations: (a) differences among autistic pupils (i.e., MAIN EFFECT of pupils), (b) differences within the coder on four occasions (i.e., main effect of occasions), (c) different observations of pupils on different occasions (i.e., INTERACTION EFFECT

between pupils and occasions), and (d) unknown errors (e.g., RANDOM ERRORS and unspecified but SYSTEMATIC ERRORS). All sources of variation are assumed random so that we can address the question, "How well does the coder's observation on one occasion generalize to other occasions?" Because the third and the fourth sources cannot be distinguished from the data, both are collectively considered as RESIDUALS. The magnitude of variation resulting from the above sources can be estimated by a VARIANCE COMPONENTS MODEL.

Based on the formulas presented in Table 2, the estimated variance components are derived and reported in Table 3. As seen, the variance component for pupils merely accounts for 4% of the total variance, which suggests that an observation based on *one* occasion can hardly distinguish among pupils' behaviors. A large proportion of variance in residuals further suggests that pupils' eye contacts fluctuate across different occasions.

To assess intracoder reliability, we first identify the appropriate measurement errors of interest, which are determined by the interpretations or decisions made by the researchers. Researchers may desire to interpret the data based on how well a pupil behaves relative to other pupils (i.e., relative decisions), or the absolute number of eye contacts (i.e., absolute decisions). Measurement error for the relative decision (σ_R^2) can be estimated by

$$\hat{\sigma}_R^2 = \frac{\hat{\sigma}_{po,e}^2}{n_o^*},$$

Table 2 Sources of Variance, Correspondent Notations, and Computation Formulas

Source of Variation	Sums of Square	Degrees of Freedom	Mean Squares	Estimated Variance Components
Pupils (<i>p</i>)	SS_p	$df_p = n_p - 1$	$MS_p = SS_p/df_p$	$\hat{\sigma}_p^2 = (MS_p - \hat{\sigma}_{po,e}^2)/n_o$
Occasions (<i>o</i>)	SS_o	$df_o = n_o - 1$	$MS_o = SS_o/df_o$	$\hat{\sigma}_o^2 = (MS_o - \hat{\sigma}_{po,e}^2)/n_p$
Residuals (<i>po, e</i>)	$SS_{po,e}$	$df_{po,e} = df_p \times df_o$	$MS_p = SS_{po,e}/df_{po,e}$	$\hat{\sigma}_{po,e}^2 = MS_{po,e}$

NOTE: n_p and n_o are number of pupils and number of occasions, respectively.

Table 3 Sources of Variance Based on Data in Table 1

Source of Variation	Sums of Square	Degrees of Freedom	Mean Squares	Estimated Variance Components	Percentage of Total Variance (%)
Pupils (<i>p</i>)	13.53	9	1.50	0.046	4
Occasions (<i>o</i>)	5.68	3	1.89	0.057	4
Residuals (<i>po, e</i>)	35.58	27	1.32	1.318	93

Table 4 Measurement Errors and Intracoder Reliability Estimates Under Different Types of Decisions and Numbers of Occasions

Occasions	Relative Decisions		Absolute Decisions	
	Measurement Error	Intracoder Reliability	Measurement Error	Intracoder Reliability
1 day	1.318	0.033	1.364	0.032
2 days	0.659	0.065	0.682	0.063
3 days	0.439	0.094	0.454	0.091
4 days	0.329	0.122	0.341	0.118
5 days	0.263	0.148	0.272	0.144
10 days	0.131	0.258	0.136	0.252
20 days	0.065	0.411	0.068	0.402
30 days	0.043	0.511	0.045	0.502
40 days	0.032	0.582	0.034	0.574
80 days	0.016	0.736	0.017	0.729
120 days	0.010	0.807	0.011	0.801
160 days	0.008	0.848	0.008	0.843

whereas that for the absolute decision (σ_A^2) can be estimated by

$$\hat{\sigma}_A^2 = \frac{\hat{\sigma}_o^2}{n_o^*} + \frac{\hat{\sigma}_{po,e}^2}{n_o^*},$$

where n_o^* refers to the number of occasions used in the study (e.g., Table 1), or the number of occasions researchers should choose to improve the observation consistency.

For both the relative decision and the absolute decision, intracoder reliability (or termed the generalizability coefficient in G theory) is estimated by

$$G_R = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_R^2}$$

and

$$G_A = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_A^2},$$

respectively. Intracoder reliability estimates under different types of decisions and numbers of occasions are presented in Table 4. As shown, the more occasions a coder observes each pupil, the fewer measurement errors there are. It would require more than 80 days of the coder's observations to reach a satisfactory level of consistency. The data also show poor intracoder reliability when the coder observes the pupils on only four occasions.

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

- Brennan, R. L. (1983). *Elements of generalizability*. Iowa City, IA: American College Testing Program.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

INVESTIGATOR EFFECTS

Investigator effects are those sources of artifact or error in scientific inquiry that derive from the investigator. It is useful to think of two major types of effects, usually unintentional, that scientists can have upon the results of their research. The first type operates, so to speak, in the mind, eye, or hand of the scientist. It operates without affecting the actual response of the human participants or animal subjects of the research. It is not interactional. The second type of investigator effect is interactional. It operates by affecting the actual response of the subject of the research.

Noninteractional effects include (a) observer effects, (b) interpreter effects, and (c) intentional effects.

(a) *Observer effects* refer to errors of observation made by scientists in the perception or recording of the events they are investigating. The analysis of a series of 21 studies involving 314 observers who recorded a total of 139,000 observations revealed that about 1% of the observations were in error, and that when errors were made, they occurred two thirds of the time in the direction of the observer's hypothesis.

(b) *Interpreter effects* refer to differences in the theoretical interpretations that different scientists give to the same set of observations. For many years, for example, investigators of the effectiveness of psychotherapy disagreed strongly with one another in the interpretation of the available studies. It was not until systematic, quantitative summaries of the hundreds of studies on the effectiveness of psychotherapy became available that this particular issue of interpreter effects became fairly well resolved.

(c) *Intentional effects* refer to instances of outright dishonesty in science. Perhaps the most common example is simple data fabrication, in which

nonoccurring but desired observations are recorded instead of the honest observations of the events purported to be under investigation.

Interactional investigator effects include (a) biosocial, (b) psychosocial, (c) situational, (d) modeling, and (e) expectancy effects.

(a) *Biosocial effects* refer to those differences in participants' responses associated with, for example, the sex, age, or ethnicity of the investigator.

(b) *Psychosocial effects* refer to those differences in participants' responses associated with, for example, the personality or social status of the investigator.

(c) *Situational effects* refer to those differences in participants' responses associated with investigator differences in such situational variables as research experience, prior acquaintance with participants, and the responses obtained from earlier-contacted participants.

(d) *Modeling effects* refer to those differences in participants' responses associated with investigator differences in how they themselves responded to the task they administer to their participants.

(e) *Expectancy effects* refer to those differences in participants' responses associated with the investigator's expectation for the type of response to be obtained from the participant. Expectations or hypotheses held by investigators in the social and behavioral sciences have been shown to affect the behavior of the investigator in such a way as to bring about the response the investigator expects from the participant. This effect, referred to most commonly as the EXPERIMENTER EXPECTANCY EFFECT, is a special case of interpersonal expectancy effects that has been found to occur in many situations beyond the laboratory.

—Robert Rosenthal

See also PYGMALION EFFECT

REFERENCES

- Rosenthal, R. (1976). *Experimenter effects in behavioral research: Enlarged edition*. New York: Irvington.
- Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart and Winston.
- Rosnow, R. L., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. New York: W. H. Freeman.

IN VIVO CODING

To understand in vivo coding, one must first understand CODING. A code is a concept, a word that signifies "what is going on in this piece of data." Coding, on the other hand, is the analytic process of examining data line by line or paragraph by paragraph (whatever is your style) for significant events, experiences, feelings, and so on, that are then denoted as concepts (Strauss & Corbin, 1998). Sometimes, the analyst names the concept. Other times, the words of the respondent(s) are so descriptive of what is going on that they become the designated concept. These descriptive words provided by respondents are what have come to be known as in vivo concepts (Glaser & Strauss, 1967).

In vivo concepts are usually snappy words that are very telling and revealing. The interesting thing about in vivo codes is that a researcher knows the minute the idea is expressed by a respondent that this is something to take note of. The term that is used expresses meaning in a way far better than any word that could be provided by the analyst (Strauss, 1987). For example, suppose a researcher were doing a study of what goes on at cocktail parties. One interviewee might say something like,

Many people hate going to cocktail parties, but I love them. When I go, I take the opportunity to *work the scene*. It may seem like idle chatter, but in reality, I am making mental notes about future business contacts, women I'd like to date, possible tennis partners, and even investment opportunities.

Working the scene is a great concept. It doesn't tell us everything that goes on at cocktail parties, but it does convey what this man is doing. And if he does this, then perhaps other people also are *working the scene*. The researcher would want to keep this concept in mind when doing future interviews to see if other people describe actions that could also be labeled as *working the scene*.

In contrast, an analyst-derived code would look something like the following. While observing at another cocktail party, the researcher notices two women talking excitedly about work projects in which they are engaged. In another area, a man and a woman are also talking about their work, and further on, another group of people are doing likewise. The

researcher labels this talking about work at a cocktail party as “social/work talk.” This term describes what is going on but it’s not nearly as snappy or interesting as *working the scene*.

Not every interview or observation yields interesting in vivo codes, but when these do come up in data, the researcher should take advantage of them. It is important that analysts are alert and sensitive to what is in the data, and often, the words used by respondents are the best way of expressing that. Coding can be a laborious and detailed process, especially when analyzing line by line. However, it is the detailed coding that often yields those treasures, the terms that we have come to know as in vivo codes.

—Juliet M. Corbin

REFERENCES

- Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Chicago: Aldine.
- Strauss, A. (1987). *Qualitative analysis for social scientists*. Cambridge, UK: Cambridge University Press.
- Strauss, A., & Corbin, J. (1998). *Basics of qualitative analysis*. Thousand Oaks, CA: Sage.

ISOMORPH

An isomorph is a theory, model, or structure that is similar, equivalent, or even identical to another. Two theories might be conceptually isomorphic, for example, but use different empirical measures and thus be empirically distinct. Or, two theories might appear to be different conceptually or empirically, but be isomorphic in their predictions.

—Michael S. Lewis-Beck

ITEM RESPONSE THEORY

DEFINITION AND APPLICATION AREAS

Tests and questionnaires consist of a number of items, denoted J , that each measure an aspect of the same underlying psychological ability, personality trait, or attitude. Item response theory (IRT) models use the data collected on the J items in a sample of N

respondents to construct scales for the measurement of the ability or trait.

The score on an item indexed j ($j = 1, \dots, J$) is represented by a random variable X that has realizations x_j . Item scores may be

- *Dichotomous*, indicating whether an answer to an item was correct (score $x_j = 1$) or incorrect (score $x_j = 0$)
- *Ordinal polytomous*, indicating the degree to which a respondent agreed with a particular statement (ordered integer scores, $x_j = 0, \dots, m$)
- *Nominal*, indicating a particular answer category chosen by the respondent, as with multiple-choice items, where one option is correct and several others are incorrect and thus have nominal measurement level
- *Continuous*, as with response times indicating the time it took to solve a problem

Properties of the J items, such as their difficulties, are estimated from the data. They are used for deciding which items to select in a paper-and-pencil test or questionnaire, and more advanced computerized measurement procedures. Item properties thus have a technical role in instrument construction and help to produce high-quality scales for the measurement of individuals.

In an IRT context, abilities, personality traits, and attitudes underlying performance on items are called latent traits. A latent trait is represented by the random variable θ , and each person i ($i = 1, \dots, N$) who takes the test measuring θ has a scale value θ_i . The main purpose of IRT is to estimate θ for each person from his or her J observed item scores. These estimated measurement values can be used to compare people with one another or with an external behavior criterion. Such comparisons form the basis for decision making about individuals.

IRT originated in the 1950s and 1960s (e.g., Birnbaum, 1968) and came to its full bloom afterwards. Important fields of application are the following:

- *Educational measurement*, where tests are used for grading examinees on school subjects, selection of students for remedial teaching and follow-up education (e.g., the entrance to university), and certification of professional workers with respect to skills and abilities

- *Psychology*, where intelligence measurement is used, for example, to diagnose children's cognitive abilities to explain learning and concentration problems in school, personality inventories to select patients for clinical treatment, and aptitude tests for job selection and placement in industry and with the government
- *Sociology*, where attitudes are measured toward abortion or rearing children in single-parent families, and also latent traits such as alienation, Machiavellianism, and religiosity
- *Political science*, where questionnaires are used to measure the preference of voters for particular politicians and parties, and also political efficacy and opinions about the government's environmental policy
- *Medical research*, where health-related quality of life is measured in patients recovering from accidents that caused enduring physical damage, radical forms of surgery, long-standing treatment using experimental medicine, or other forms of therapy that seriously affect patients' experience of everyday life
- *Marketing research*, where consumers' preferences for products and brands are measured

Each of these applications requires a quantitative scale for measuring people's proficiency, and this is what IRT provides.

This entry first introduces the assumptions common to IRT models. Then, several useful distinctions are made to classify different IRT models. Finally, some useful applications of IRT models are discussed.

COMMON ASSUMPTIONS OF IRT MODELS

Dimensionality of Measurement

The first assumption enumerates the parameters necessary to summarize the performance of a group of individuals that takes the test or questionnaire. For example, if a test is assumed to measure the ability of spatial orientation, as in subtests of some intelligence test batteries, an IRT model that assumes only one person parameter may be used to describe the data. This person parameter represents for each individual his or her level of spatial orientation ability. An IRT model with one person parameter is a strictly unidimensional (UD) model. Alternatively, items in another test may measure a mixture of arithmetic ability and word comprehension, for example, when

arithmetic exercises are embedded in short stories. Here, a two-dimensional IRT model may be needed. Other test performances may be even more complex, necessitating multidimensional IRT models to explain the data structure. For example, in the arithmetic example, some items may also require general knowledge about stores and the products sold there (e.g., when calculating the amount of money returned at the cash desk), and others may require geographical knowledge (e.g., when calculating the distance between cities). Essentially unidimensional models assume all the items in a test to measure one dominant latent trait and a number of nuisance traits that do not disturb measurement of the dominant θ when tests are long. For example, a personality inventory on introversion may also measure anxiety, social intelligence, and verbal comprehension, each measured only weakly by one or two items and dominated by introversion as the driving force of responses.

Relationships Between Items

Most IRT models assume that, given the knowledge of a person's position on the latent trait or traits, the joint distribution of his or her item scores can be reconstructed from the marginal frequency distributions of the J items. Define a vector $\mathbf{X} = (X_1, \dots, X_J)$ with realization $\mathbf{x} = (x_1, \dots, x_J)$ and let $\boldsymbol{\theta}$ be the vector with latent traits needed to account for the test performance of the respondents. In IRT, the marginal independence property is known as LOCAL INDEPENDENCE (LI), and defined as

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^J P(X_j = x_j|\boldsymbol{\theta}). \quad (1)$$

At the behavior level, LI means that during test administration, each item is answered independent of the other items. There is no learning or development due to practice while the respondent tries to solve the items; thus, the measurement procedure does not influence the measurement outcome. However, improving students' ability through practice may be the purpose of a test program, as is typical of dynamic testing. Then, an IRT model may assume that practice produces successively more skills, each represented by a new latent trait. This means that possible dependencies between items during testing not explained by $\boldsymbol{\theta}$ are absorbed by simply adding more θ s.

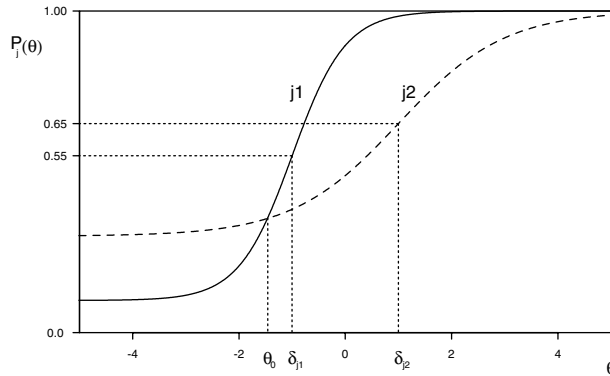


Figure 1 Two IRFs Under the Three-Parameter Logistic Model (Parameter Values: $\gamma_{j1} = 0.10$, $\gamma_{j2} = 0.30$; $\delta_{j1} = -1.00$, $\delta_{j2} = 1.00$; $\alpha_{j1} = 2.00$, $\alpha_{j2} = 1.00$); Intersection Point at $\theta_0 = 1.46$; $P(\theta_0) = 0.36$

Relationships Between Items and the Latent Trait

For unidimensional dichotomous items, the item response function (IRF) describes the relationship between item score X_j and latent trait θ , and is denoted $P_j(\theta) = P(X_j = 1|\theta)$. Figure 1 shows two typical monotone increasing IRFs (assumption M) from the three-parameter logistic model that is introduced shortly. Assumption M formalizes that a higher θ drives the response process such that a correct answer to the item becomes more likely. The two IRFs differ in three respects, however.

First, the IRFs have different slopes. A steeper IRF represents a stronger relationship between the item and the latent trait. IRT models have parameters related to the steepest slope of the IRF. This slope parameter, denoted α_j for item j , may be compared with a REGRESSION COEFFICIENT in a LOGISTIC REGRESSION model. Item j_1 (solid curve) has a steeper slope than Item j_2 (dashed curve); thus, it has a stronger relationship with θ and, consequently, discriminates better between low and high values of θ .

Second, the locations of the IRFs are different. Each IRF is located on the θ scale by means of a parameter, δ_j , that gives the value of θ where the IRF is halfway between the lowest and the highest conditional probability possible. Item j_1 (solid curve) is located more to the left on the θ scale than Item j_2 (dashed curve), as is shown by their location parameters: $\delta_{j1} < \delta_{j2}$. Because the slopes are different, the IRFs intersect. Consequently, for the θ s to the left of the intersection

point, θ_0 , Item j_1 is less likely to be answered correctly than Item j_2 . Thus, for these θ s, Item j_1 is more difficult than Item j_2 . For the θ s to the right of θ_0 , the item difficulty ordering is reversed.

Third, the IRFs have different lower asymptotes, denoted γ_j . Item parameter γ_j is the probability of a correct answer by people who have very low θ s. In Figure 1, Item j_1 has the lower γ parameter. This parameter is relevant, in particular, for multiple-choice items where low- θ people often guess with nonzero probability for the correct answer. Thus, multiple-choice Item j_2 is more liable to guessing than Item j_1 .

Three-Parameter Logistic Model. The IRFs in Figure 1 are defined by the three-parameter logistic model. This model is based on assumptions UD and LI, and defines the IRF by means of a logistic function with the three item parameters discussed:

$$P_j(\theta) = \gamma_j + (1 - \gamma_j) \frac{\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}. \quad (2)$$

One-Parameter Logistic Model. Another well-known IRT model is the one-parameter logistic model or Rasch model, also based on UD and LI, that assumes that for all J items in the test, $\gamma = 0$ and $\alpha = 1$ (the value 1 is arbitrary; what counts is that α_j is a constant, a , for all j). Thus, this model (a) is not suited for fitting data from multiple-choice items because that would result in positive γ s; (b) assumes equally strong relationships between all item scores and the latent trait ($\alpha_j = a$ for all j); and (c) allows items to differ only in difficulty δ_j , $j = 1, \dots, J$.

Linear Logistic Multidimensional Model. Figure 2 shows an IRF as a three-dimensional surface in which the response probability depends on two latent traits, $\theta = (\theta_1, \theta_2)$. The slope of the surface is steeper in the θ_2 direction than in the θ_1 direction. This means that the probability of having the item correct depends more on θ_2 than θ_1 . Because θ_2 matters more to a correct answer than θ_1 , by using this item for measurement, people are better distinguished on the θ_2 scale than on the θ_1 scale. The two slopes may be different for other items that measure the composite, θ . The location (not visible in Figure 2) of this item is related (but not identical) to the distance of the origin of the space to the point of steepest slope in the direction from the origin. The γ_j parameter represents the probability of a correct answer when both θ s are very low.

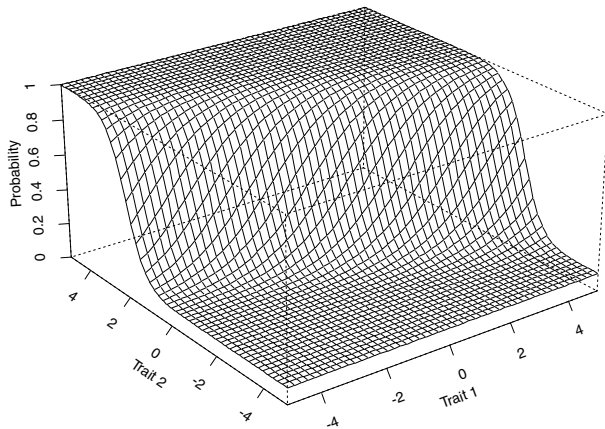


Figure 2 Item Response Surface Under the Linear Logistic Multidimensional Model (Parameter Values: $\gamma_j = 0.10$; $\delta_j = 0.00$; $\alpha_{j1} = 1.00$; $\alpha_{j2} = 2.50$)

The item response surface in Figure 2 originated from Reckase's linear logistic multidimensional model (Van der Linden & Hambleton, 1997, pp. 271–286). This IRT model has a multidimensional θ , and slope parameters collected in a vector, α . The item response surface is given by

$$P_j(X_j = 1|\theta) = \gamma_j + (1 - \gamma_j) \frac{\exp(\alpha'_j \theta + \delta_j)}{1 + \exp(\alpha'_j \theta + \delta_j)}. \quad (3)$$

Graded Response Model. Finally, Figure 3 shows the item step response functions (ISRFs) (solid curves) of an item under Samejima's graded response model (Van der Linden & Hambleton, 1997, pp. 85–100) for polytomous item scores. For ordered integer item scores ($x_j = 0, \dots, m$) and a unidimensional latent trait, each conditional response probability $P(X_j \geq x_j|\theta)$, is modeled separately by logistic functions. These functions have a location parameter that varies between the item's ISRFs and a constant slope parameter. Between items, slope parameters are different. It may be noted that for $x_j = 0$, the ISRF equals 1, and that for a fixed item, the other m ISRFs cannot intersect by definition. ISRFs of different items can intersect, however, due to different slope parameters (see Figure 3). Polytomous IRT models are more complex mathematically than dichotomous IRT models because they involve more response functions and a greater number of item parameters. Many other IRT models exist; see Van der Linden and Hambleton (1997) for an extensive overview.

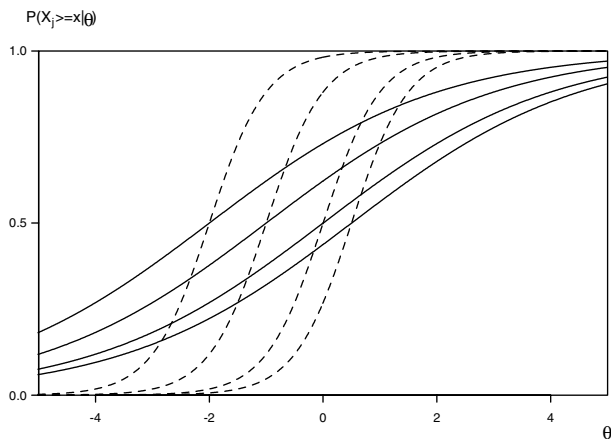


Figure 3 ISRFs of Two Items With Five Answer Categories Each, Under the Graded Response Model (No Parameter Values Given)

Nonparametric Models. The models discussed so far are parametric models because their IRFs or ISRFs are parametric functions of θ . Nonparametric IRT models put only order restrictions on the IRFs but refrain from a more restrictive parametric definition. This is done in an attempt to define measurement models that do not unduly restrict the test data structure, while still imposing enough structure to have ordinal measurement of people on the θ scale. This ordering is estimated by means of the sum of the item scores, which replaces θ as a summary of test performance. Thus, flexibility is gained at the expense of convenient mathematical properties of parametric—in particular, logistic—functions. See Boomsma, Van Duijn, and Snijders (2001) for discussions of parametric and nonparametric IRT models, and Sijtsma and Molenaar (2002) for an introduction to nonparametric IRT models.

APPLICATIONS OF IRT MODELS

Test Construction, Calibration. IRT models imply the scale of measurement, which is either ordinal or interval. When the IRT model fits the data, the scale is assumed to hold for the particular test. Scales are calibrated and then transformed so as to allow a convenient interpretation. For example, the θ metric can be replaced by the well-known total score or percentile scales. The scale may constitute the basis for individual diagnosis, as in educational and psychological measurement, and for scientific research.

Equating and Item Banks, Adaptive Testing. The θ metric is convenient for the equating of scales based on different tests for the same latent trait, with the purpose of making the measurements of pupils who took these different tests directly comparable. Equating may be used to construct an item bank, consisting of hundreds of items that measure the same latent trait, but with varying difficulty and other item properties. New tests can be assembled from an item bank. Tests for individuals can be assembled by selecting items one by one from the item bank so as to reduce measurement error in the estimated θ as quickly as possible. This is known as adaptive testing. It is convenient in large-scale testing programs in education, job selection, and placement.

Differential Item Functioning. People with different backgrounds are often assessed with the same measurement instruments. An important issue is whether people having the same θ level, but differing on, for example, gender, socioeconomic background, or ethnicity, have the same response probabilities on the items from the test. If not, the test is said to exhibit differential item functioning. This can be investigated with IRT methods. Items functioning differently between groups may be replaced by items that function identically.

Person-Fit Analysis. Respondents may be confused by the item format; they may be afraid of situations, including a test, in which they are evaluated; or they may underestimate the level of the test and miss the depth of several of the questions. Each of these

mechanisms, as well as several others, may produce a pattern of J item scores that is unexpected given the predictions from an IRT model. Person-fit methods have been proposed to identify nonfitting item score patterns. They may contribute to the diagnosis of the behavior that caused the unusual pattern.

Cognitive IRT Models. Finally, cognitive modeling has taken measurement beyond assigning scores to people, in that the cognitive process or the solution strategy that produced these scores is part of the IRT model, and measurement is related to a psychological explanation. This approach may lead to the identification of skills that are insufficiently mastered or, at the theoretical level, to a better understanding of the processes underlying test performance.

—Klaas Sijtsma

REFERENCES

- Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Boomsma, A., Van Duijn, M. A. J., & Snijders, T. A. B. (Eds.). (2001). *Essays on item response theory*. New York: Springer.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to non-parametric item response theory*. Thousand Oaks, CA: Sage.
- Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

J

JACKKNIFE METHOD

The jackknife method is a resampling procedure that can be used to generate an empirical SAMPLING DISTRIBUTION. Such distributions can be used to run statistical hypothesis tests, construct CONFIDENCE INTERVALS, and estimate STANDARD ERRORS. The jackknife method is parallel to the BOOTSTRAP and the permutation (or randomization) methods.

Rodgers (1999) presented a TAXONOMY organizing these methods based on size of the resampled samples and whether sampling occurs with or without replacement. The jackknife method involves repeatedly drawing samples without replacement that are smaller than the original sample. Originally, the jackknife was defined as an approach in which one or more data points are randomly deleted, a conceptualization equivalent to randomly sampling without replacement.

The jackknife method was originally developed by Quenouille (1949) and Tukey (1958); Efron (1982), developer of the bootstrap, named it the “Quenouille-Tukey jackknife.” The method was summarized by Mosteller and Tukey (1977): “The basic idea is to assess the effect of each of the groups into which the data have been divided... through the effect upon the body of data that results from *omitting* that group” (p. 133). The jackknife has broadened, as implied in Rodgers (1999): “Let the observed data define a population of scores. Randomly sample from this population *without replacement* to fill the groups with a *smaller number of scores than in the original sample*” (p. 46).

In general, an empirical sampling distribution can be generated by deleting one, two, or a whole group of observations; computing a relevant statistic within the new sample; and generating a distribution of these statistics by repeating this process many times.

As an example, consider a research design with RANDOM ASSIGNMENT of 10 participants to a treatment or a control group ($n = 5$ per group). Did the treatment result in a reliable increase? Under the NULL HYPOTHESIS of no group difference, the 10 scores are sorted into one or the other group purely by chance. The usual formal evaluation of the null hypothesis involves computation of a two-independent group t -statistic measuring a standardized mean difference between the two groups, which is then evaluated in relation to the presumed distribution under the null hypothesis, modeled by the theoretical t -distribution. This evaluation rests on a number of parametric statistical assumptions, including normality and independence of errors and homogeneity of variance. Alternatively, the null distribution can be defined as an empirical t -distribution using resampling, a process requiring fewer assumptions.

In the jackknife method, we might decide to delete 1 observation from each group (or, equivalently, to sample without replacement from the original 10 observations until each group has 4 observations). There are 3,150 possible resamples. The t -statistic is computed in each new sample, resulting in a distribution of t s approximating the population t -distribution under the null hypothesis. The original t -statistic is evaluated in relation to this empirical sampling distribution to draw

a conclusion about group differences. Furthermore, a standard error of the t -statistic can be estimated from the standard deviation of the resampled t -statistics, and confidence intervals can also be constructed.

—Joseph Lee Rodgers

REFERENCES

- Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.
- Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, 11, 18–84.
- Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, 34, 441–456.
- Tukey, J. W. (1958). Bias and confidence in not quite large samples [abstract]. *Annals of Mathematical Statistics*, 29, 614.

K

KEY INFORMANT

Also referred to as a key consultant (Werner & Schoepfle, 1987), a key informant is an individual who provides in-depth, expert information on elements of a culture, society, or social scene. Such in-depth information tends to be gathered in repeated sets of SEMI-STRUCTURED and UNSTRUCTURED INTERVIEWS in mostly natural, informal settings (e.g., the key informant's home). Aside from the depth and breadth of information, what sets a key informant apart from other types of interviewees is the duration and intensity of the relationship between the researcher and the informant. Three primary types of interviewees in social research include subjects, respondents, and informants. Subjects are generally individuals interviewed in the course of a social experiment. Respondents are individuals who are interviewed in the course of a social SURVEY. Finally, informants are individuals who are interviewed in a more in-depth, less structured manner (semi-structured, unstructured interviews), most often in a field setting. In any given study, individuals can move from the role of an informant to respondent to subject depending on the interviewing context (e.g., survey, IN-DEPTH INTERVIEW). Key informants are a subtype of informant in which a special relationship exists between interviewer (e.g., ethnographer) and interviewee.

Key informants are selected on the basis of an informant's characteristics, knowledge, and rapport with the researcher. Johnson (1990) describes two primary criteria for selecting informants. The first set of criteria concerns the form and type of expert knowledge and the

theoretical representativeness of the informant, where such representativeness refers to the characteristics of the informant with respect to theoretically important factors (e.g., status, position in an organization) or the emergent features of some set of data (e.g., the position of the informant in the distribution of incomes). The second set of criteria concerns the informant's personality, interpersonal skills, and the chemistry between the researcher and the informant. Whereas the first set of criteria places the informant within the framework of the study's goals and objectives, the second set of criteria reflects the important nature of the collaborative relationship between the researcher and key informant. Once the researcher understands the place of an individual within the realm of theoretical criteria, key informant selection can then be based on the potential for rapport, thus limiting potential biases that may arise from selection on the basis of personal characteristics alone.

What further separates key informants from the other main types of interviewees is their potential involvement throughout the various phases of the research enterprise. Key informants and their expert knowledge and advice not only are a source of data but also can be used to develop more valid formal survey instruments (e.g., help in better framing of survey questions), provide running commentaries on cultural events or scenes, aid in the interpretation of study results from more systematically collected data sources (e.g., survey results), and help in determining the VALIDITY and RELIABILITY of data (e.g., genealogies), to mention a few. Key informants are an essential component of any ethnographic enterprise.

—Jeffrey C. Johnson

REFERENCES

- Johnson, J. C. (1990). *Selecting ethnographic informants*. Newbury Park, CA: Sage.
- Werner, O., & Schoepfle, G. M. (1987). *Systematic fieldwork* (2 vols.). Newbury Park, CA: Sage.

KISH GRID

The Kish grid was developed by Leslie Kish in 1949 and is commonly used by those conducting large-scale sample SURVEYS. It is a technique used in equal-probability SAMPLING for selecting cases at random when more than one case is found to be eligible for inclusion when the interviewer calls at a sampled address or household. For example, a sample of addresses may be drawn from a list of all postal addresses in a given area, and the sample design assumes that each address on that list will contain only one household. When the interviewer calls at the sampled address, he or she may find more than one household at the address and will need to select one at random (or whatever number is specified in the sample design) to include in the sample. Most important, this must be done systematically and in a way that provides a known probability of selection for every household included.

The same problem occurs when selecting an individual within a household. A household may consist of one person only or many persons. Households that contain more than one member of the population become clusters of population elements, so they must be translated from a sample of households to a sample of individuals. If the survey design asks for a specific type of individual to be interviewed (e.g., the person responsible for the accommodation or for all individuals age 16 years or older to be included), the problem of selection does not arise. However, if the survey design asks for one individual to be included from each household, the Kish grid is used to randomly select the individual for inclusion in the sample from each selected household.

The Kish grid is a simple procedure that is easy for interviewers to implement in the field. It provides a clear and consistent procedure for randomly selecting cases for inclusion in the sample with a known probability of selection. The interviewer first needs an objective way to order the members of the household or number of dwellings before they can be selected. To select an individual within a household, the interviewer has a pre-prepared grid that asks him or her to list all the

individuals living within the household systematically, usually according to their age and sex, and each person is given a serial number. Who is selected for interview depends on the total number of persons within the household. The grid includes a selection table that tells the interviewer which numbered individual to select. The interviewer simply reads across the table from a column listing the number of persons in the household to the serial number he or she should select for interview. The Kish grid is a standard technique that is widely used and can be adapted to meet the needs of different survey designs.

—Heather Laurie

REFERENCES

- Kish, L. (1949). A procedure for objective respondent selection within a household. *Journal of the American Sociological Association*, 44, 380–387.
- Kish, L. (1965). *Survey sampling*. New York: John Wiley.

KOLMOGOROV-SMIRNOV TEST

There are two versions, the one-sample GOODNESS-OF-FIT test and the two-sample test.

ONE-SAMPLE TEST

Several alternatives for one-sample tests involving ordinal data are available (Gibbons & Chakraborti, 1992; Siegel & Castellan, 1988) with advantages that vary slightly with the situation. The Kolmogorov-Smirnov one-sample test is purported to be more powerful than the CHI-SQUARE goodness-of-fit test for the same data, particularly when the sample is small. The test is appropriate for ranked data, although it is truly a test for the match of the distribution of sample data to that of the population; consequently, the distributions can include those other than the normal distribution for populations. It also assumes an underlying continuous distribution, even though the data may be in unequal steps (see Gibbons & Chakraborti, 1992), which leads to alternative calculations, as will be seen below.

The test itself compares the *cumulative* distribution of the data with that for the expected population distribution. It assumes an underlying continuous distribution and has a definite advantage over the CHI-SQUARE TEST when the sample size is small, due

Table 1 Data for a Kolmogorov-Smirnov One-Sample Test for Ranked Data Where the Population Distribution Is a Mathematical Function

<i>School Salary</i>	f_o	CF_o School	CF_e County	$CF_{oi} - F_{ei}$ Adjacent Rows	$CF_{o(i-1)} - CF_{ei}$ Offset Rows
9,278	2	0.100	0.131	-0.031	-0.131
10,230	1	0.150	0.168	-0.018	-0.068
11,321	1	0.200	0.218	-0.018	-0.068
13,300	1	0.250	0.326	-0.076	-0.126
14,825	2	0.350	0.422	-0.072	-0.172
16,532	2	0.450	0.535	-0.085	-0.185
18,545	1	0.500	0.664	-0.164	-0.214
21,139	3	0.650	0.804	-0.154	-0.304
22,145	2	0.750	0.847	-0.097	-0.197
24,333	2	0.850	0.918	-0.068	-0.168
28,775	1	0.900	0.983	-0.083	-0.133
30,987	1	0.950	0.994	-0.044	-0.094
32,444	1	1.000	0.997	0.003	-0.047
Total	20				

to the limitation of the minimum expected value. The maximum magnitude of the divergence between the two distributions is the test factor and is determined simply by the largest absolute difference between cumulative *relative* frequencies, the fractions of the whole. This is checked against a table of maximum differences expected to be found for various α .

The difficulty in appropriately applying this test lies in the nature of the intervals for the ranks. There are basically two types of problems illustrated in the literature, each resulting in different applications of the Kolmogorov-Smirnov test. Each of these is presented as *the* way of applying the test by some texts:

- The intervals for the ranks are not well defined, and the underlying distribution is *assumed*; therefore, the test is mathematically determined by a continuous function, for example, a normal or even a flat distribution (see Gibbons & Chakraborti, 1992). Here the sample data are stepped and the population “data” are continuous.

- The intervals are well defined and mutually exclusive, and population data exist for them by interval; thus, the population data are stepped, as well as that for the sample (see Siegel & Castellan, 1988). This makes the test equivalent to the two-sample test in which both samples are stepped.

Some examples will be used to clarify this issue.

Continuous Population Data From a Mathematical Distribution

For example, the following question is asked: Is the distribution of a particular school’s salaries typical of those across the county? Here it is assumed that the county distribution of salaries is normal. Table 1 shows the frequency of salaries in the school and the corresponding cumulative relative frequencies for these sample data, in addition to providing a cumulative relative distribution for the population based on a normal distribution. This is shown graphically in Figure 1. It is apparent that there are two possible points of comparison for each data point: the interval above and the interval below. This is shown by the two arrows for one point on the continuous population distribution, which could be considered to be matched with either of the steps below for the sample data.

Therefore, the Kolmogorov-Smirnov test requires the calculation of both differences for all points, the last two columns in Table 1. Then the maximum difference from the complete set is compared to that for the appropriate corresponding α value. The critical D_{critical} can be closely estimated from Table 2.

For this case, if α is chosen to be 0.05, then the maximum allowed is

$$D_{\text{critical}} = \frac{1.36}{\sqrt{20 + 1}} = 0.297.$$

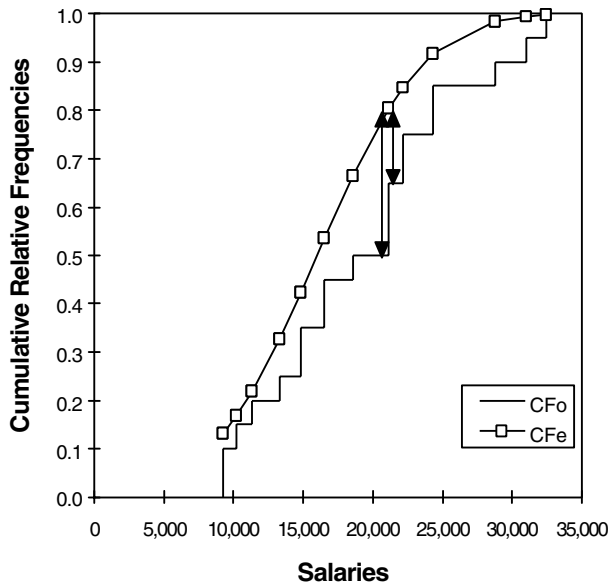


Figure 1 Observed and Expected Cumulative Relative Frequency Distributions

NOTE: These frequencies refer to a situation in which the population distribution is assumed and generated from a mathematical function (here a normal distribution).

Table 2 Critical Maximum Differences, D_{critical} , for the Kolmogorov-Smirnov Test

Sample Size	α		
	.10	.05	.01
$25 > n > 10$	$\frac{1.22}{\sqrt{n+1}}$	$\frac{1.36}{\sqrt{n+1}}$	$\frac{1.63}{\sqrt{n+1}}$
$n \geq 25$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Now reexamining Table 1, it can be seen from the last two columns of the differences that 0.305 exceeds this value, and there is a significant difference between the school's distribution and that assumed by the county. A comparison of this test with the CHI-SQUARE and T -TEST on the same data can be found in Black (1999). The point to note is that the three tests for a given set of data not only provide different levels of *power* but also answer essentially three different questions.

Stepped Population Data From a Full Survey

In this situation, the sample data are compared to actual population data, whereby the frequencies for

each of the ordinal categories for both are known. To illustrate this, let us use the sample data as before but compare it to a county register of teachers. The data and corresponding stepped graphs are shown in Figure 2, where it becomes obvious that only data within intervals are to be compared. For example, the double-ended arrow shows the difference between one pair of intervals. Therefore, the Kolmogorov-Smirnov one-sample test will only use the adjacent differences in cumulative relative frequencies (see Table 3), and the largest difference will be compared with the minimum for the chosen α . The largest difference is 0.278, which is less than the 0.297 found above (the sign does not make any difference here), and therefore there is no significant difference.

Such a distribution-free comparison will have advantages in many situations when data do not fit a predetermined distribution. To carry out the equivalent test in SPSS, one must use the two-sample function, in which one of the sets of data is that of the population.

TESTS FOR TWO INDEPENDENT SAMPLES

Not always will designs of studies involving the comparison of two groups produce interval/ratio data and lend the resolution of the hypotheses to a t -test. The

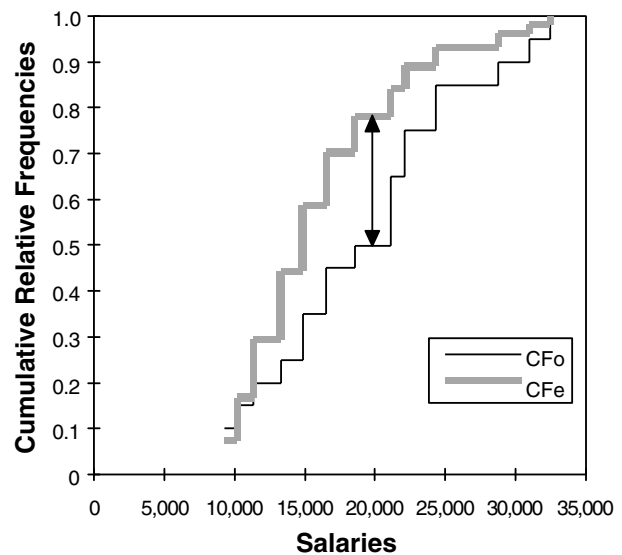


Figure 2 Observed and Expected Cumulative Relative Frequency Distributions

NOTE: These distributions correspond to the sample and the population when both sets of data are stepped (i.e., population data are from a real survey).

Table 3 Data for a Kolmogorov-Smirnov One-Sample Test for Ranked Data Where the Population Distribution Based on Ordinal Population Data

<i>School Salary</i>	f_o <i>School</i>	f_o <i>County</i>	CF_o <i>School</i>	CF_e <i>County</i>	$CF_{oi} - F_{ei}$ <i>Adjacent Rows</i>
9,278	2	260	0.100	0.097	0.003
10,230	1	287	0.150	0.205	-0.055
11,321	1	320	0.200	0.325	-0.125
13,300	1	350	0.250	0.456	-0.206
14,825	2	365	0.350	0.593	-0.243
16,532	2	287	0.450	0.700	-0.250
18,545	1	208	0.500	0.778	-0.278
21,139	3	145	0.650	0.832	-0.182
22,145	2	125	0.750	0.879	-0.129
24,333	2	100	0.850	0.916	-0.066
28,775	1	85	0.900	0.948	-0.048
30,987	1	74	0.950	0.976	-0.026
32,444	1	64	1.000	1.000	0.000
Total	20	2,670			

null hypothesis, H_0 , is basically that the nature of the frequency distributions across the categories or levels of the dependent variable is the same for both groups. In other words, does the differentiating characteristic of the two groups (INDEPENDENT VARIABLE) influence the nature of the distribution of the levels or categories of the DEPENDENT VARIABLE?

The Kolmogorov-Smirnov test checks to see if the two samples have been drawn from the same population, again with the assumption of the underlying continuous distribution. The process is the same as above, except in this case, one is comparing adjacent cumulative relative frequencies. This also requires a set of maximums that can be calculated when $n > 25$ and $m > 25$. Then,

$$D(.05) = 1.36 \sqrt{\frac{m+n}{mn}}. \quad (1)$$

For small samples, special tables are needed (see, e.g., Siegel & Castellan, 1988). Alternatively, the probabilities are calculated automatically when using SPSS. Siegel and Castellan (1988) maintain that this test is "slightly more efficient" than the WILCOXON and MANN-WHITNEY U TESTS, but when the samples are large, the reverse tends to be true.

—Thomas R. Black

REFERENCES

Black, T. R. (1999). *Doing quantitative research in the social sciences: An integrated approach to research design, measurement, and statistics*. London: Sage.

Gibbons, J. D., & Chakraborti, S. (1992). *Nonparametric statistical inference* (3rd ed.). New York: Marcel Dekker.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

KRUSKAL-WALLIS H TEST

Also known as the Kruskal-Wallis one-way analysis of variance by ranked data, this test determines whether three or more independent samples belong to the same POPULATION based on MEDIANs. It is analogous to the parametric ONE-WAY ANALYSIS OF VARIANCE (ANOVA). The MANN-WHITNEY U TEST covers the equivalent two-sample situation. The NULL HYPOTHESIS is that the medians are the same as that for the population, with the statistical test checking to see if they are close enough, allowing for natural variation in samples. The data are assumed to be at least ordinal, and there is a single underlying continuous distribution (Siegel & Castellan, 1988).

As with other tests based on rank, this one requires that all the raw data from all the samples be placed in order and assigned a rank, starting with 1 for the smallest, going to N , the number of independent observations for all k samples. The sum of all the ranks for each sample is found and the mean rank for each group is calculated. The assumption is that if the samples are really from the same population with a common mean, then these average ranks should be much the same,

Table 1 Scores for Ratings of Creepies Breakfast Cereal by Three Groups

<i>Men</i>		<i>Women</i>		<i>Children</i>	
<i>Score</i>	<i>Rank</i>	<i>Score</i>	<i>Rank</i>	<i>Score</i>	<i>Rank</i>
12	8.5	13	6.5	20	1
10	12.5	12	8.5	19	2
7	16.5	11	10.5	18	3.5
7	16.5	10	12.5	18	3.5
6	18.5	9	14.5	16	5
5	20.5	6	18.5	13	6.5
4	23	5	20.5	11	10.5
4	23	4	23	9	14.5
R_i	139		114.5		46.5
n_i	8		8		8
R_i^2/n_i	2415.13		1638.78		270.28
R_i/n_i	17.38		14.31		5.81
			Men-Children	Women-Children	Men-Women
$\alpha/k(k-1)$	0.0083	$\Delta(R/n)$	11.56	8.50	3.06
$z(0.0083)$	2.394	$\Delta(R/n)$ critical	8.46	8.46	8.46

NOTE: Tied ranks are the average of the sequential ranks. $N = 24$; $H = 11.48$; $\alpha = 0.05$; $\chi^2(2) = 5.99$.

allowing for natural variation. This is best seen in the following expression of the Kruskal-Wallis statistic, which reminds one of the calculation of variance:

$$H = \frac{12 \sum_{j=1}^k n_j (\bar{R}_j - \bar{R})^2}{N(N+1)}.$$

The statistic is also calculated by a form of the equation that makes it easy to use with raw ranks (Howell, 2002), especially on a spreadsheet (Black, 1999):

$$H = \left[\frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} \right] - 3(N+1),$$

where

R_j = sum of ranks in group j , where $j = 1$ to k groups;

$\bar{R}_j$ = average of ranks in group j ;

$\bar{R}$ = the grand mean of all ranks;

n_j = number in group j ;

N = total subjects in all groups.

When the number of groups is greater than three and the number in each group is greater than four (or even when it is three but the sample is large), then the sampling distribution is close to the CHI-SQUARE DISTRIBUTION. Also, there is a correction that can be

applied if there are a large number of ties (more than 25% of the samples), which increases the size of the H statistic (see Siegel & Castellan, 1988). If *all* values of R_i were equal to each other, then H would be 0.

For example, the data in Table 1 show scores on the trial of a new breakfast cereal, Creepies, chocolate-flavored oat cereal shaped like insects. Each person rated the cereal on four 5-point scales: texture, taste, color, and the shapes. Thus, possible scores ranged from 4 to 20. It was suspected that adults would not like them as much as children. The sample was taken so that children and adults did not come from the same families. First, the scores were rank ordered, and then H was calculated as

$$\begin{aligned} H &= \left[\frac{12}{24(24+1)} \sum_{j=1}^3 \frac{R_j^2}{n_j} \right] - 3(24+1) \\ &= \left[\frac{12}{600} (4324.2) \right] - 75 = 11.48. \end{aligned}$$

When compared to the critical value $\chi^2(df = 2) = 5.99$, it is seen that there is a difference across the three groups.

POST HOC ANALYSIS FOR KRUSKAL-WALLIS

Carrying out pairwise comparisons is simply a matter of testing the differences between pairs of averages of ranks against a standard that has been

adjusted to take into account that these are not independent. When the samples are large, their differences are approximately normally distributed (Siegel & Castellan, 1988), and the CRITICAL VALUE is found by

$$\Delta(R_i/n_i)_{\text{crit}} = z(\alpha/k(k-1)) \sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_a} + \frac{1}{n_b} \right)}.$$

Assuming that all the sample sizes are the same, this calculation has to be carried out only once. There is an adjustment that can be made if the comparisons are to be done only between each treatment and a control but not between treatments (analogous to Dunnett's post hoc test for ANOVA) (see Siegel & Castellan, 1988). For our example above,

$$\begin{aligned} \Delta(R_i/n_i)_{\text{crit}} &= z(.05/3(3-1)) \sqrt{\frac{24(24+1)}{12} \left(\frac{1}{8} + \frac{1}{8} \right)} \\ &= 2.394\sqrt{12.5} = 8.46. \end{aligned}$$

From Table 1, it can be seen that the differences between men and children, as well as between women and children, are greater than 8.46, and thus these differences are significant. There is no significant difference between men and women. Therefore, the suspicions were confirmed.

—Thomas R. Black

REFERENCES

- Black, T. R. (1999). *Doing quantitative research in the social sciences: An integrated approach to research design, measurement, and statistics*. London: Sage.
- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.
- Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

KURTOSIS

Kurtosis, skewness, and their standard errors are common univariate descriptive statistics that measure the shape of the distribution. Kurtosis is the peakedness

of a distribution, measuring the relationship between a distribution's tails and its most numerous values. Kurtosis is commonly used to screen data for NORMAL DISTRIBUTION. Values of skewness, which measures the symmetry of the distribution, and kurtosis, which measures peakedness, are generally interpreted together. By statistical convention, skewness and kurtosis both should fall in the range from +2 to -2 if data are normally distributed. The mathematical formula is as follows:

$$\text{kurtosis} = [\Sigma(X - \mu)^4 / (N\sigma^4)] - 3,$$

where

X = individual score in the distribution,

$\bar{X}$ = mean of the sample,

N = number of data points,

σ = square root of variance.

Mathematically, kurtosis is the fourth-ORDER central MOMENT. The kurtosis of a standard normal distribution is 3. By subtracting 3 from kurtosis in the formula, zero kurtosis becomes indicative of a perfect normal distribution.

It is important to obtain an adequate sample size when analyzing kurtosis as the size and even direction of kurtosis may be unstable for small sample sizes. When data are found not to be normally distributed, the researcher may wish to apply some data TRANSFORMATION to produce a normal distribution. Statistical programs typically output histograms, BOXPLOTS, box and whisker plots, stem plots, or STEM-AND-LEAF plots to display a distribution's direction and size of skewness and peakedness. These plots are useful in identifying which transformation technique to use if the data are not normally distributed. As a simple example, a normal curve may be superimposed on a histogram of responses to assess a distribution's shape. Log and square root transformations are the most common procedures for correcting positive skewness and kurtosis, whereas square root transformations correct negative skewness and kurtosis. Cubed root and negative reciprocal root transformations are used to correct extreme negative and positive skew, respectively (Newton & Rudestam, 1999).

Data may be leptokurtic, platykurtic, or mesokurtic depending on the shape of the distribution (Newton & Rudestam, 1999). Large positive values are leptokurtic and illustrate a very peaked distribution, with data clustered around the mean and with thicker or longer tails,

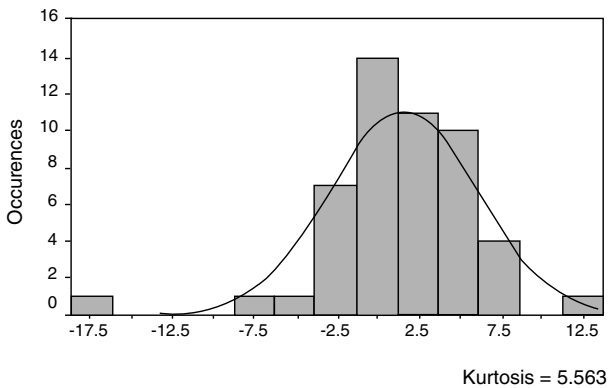


Figure 1 Leptokurtic Data

meaning there are fewer cases in the tails. As shown in Figure 1, which was obtained from hypothetical data, leptokurtic data clustered around the mean and with fewer cases in the tails, in comparison to normally distributed data, resulted in moderate positive kurtosis.

Positive skewness and peakedness indicate that the mean is larger than the median. Large negative values are platykurtic, with histograms indicating a very flat distribution with less clustering around the mean and lighter or shorter tails, meaning there are more cases in the tails than would appear in a normal distribution. Negative skewness and peakedness indicate that the median is larger than the mean. Mesokurtic data approximate a normal distribution, which may be either slightly negative or slightly positive. Low skewness and peakedness reflect relative closeness of the mean and median.

—Amna C. Cameron

REFERENCES

Newton, R. R., & Rudestam, K. E. (1999). *Your statistical consultant: Answers to your data analysis questions*. Thousand Oaks, CA: Sage.

L

LABORATORY EXPERIMENT

All EXPERIMENTS involve random assignment of participants to the conditions of the study, an INDEPENDENT VARIABLE (IV) and a DEPENDENT VARIABLE (DV). Experiments may be conducted in highly constrained laboratory settings or in field contexts, which generally are characterized by somewhat less control. Both venues offer advantages and disadvantages. In the laboratory, the researcher typically has control over all external forces that might affect the participant—the degree of control over the setting is extreme. In field experiments, this often is not so; hence, conclusions are sometimes more tentative in such contexts, owing to the possibility of the effect of uncontrolled variations.

Random assignment in experimentation means that from a predefined pool, all participants have an equal chance of being assigned to one or another condition of the study. After random assignment, participants assigned to different conditions receive different levels of the IV. In some designs, participants will have been pretested (i.e., assessed on a measure that theoretically is sensitive to variations caused by different levels of the IV). In other designs, only a posttreatment DV is used (see Crano & Brewer, 2002). After the treatment application, they are assessed again, as indicated in Figure 1, which shows two groups, a treated group and a control group (which did not receive the experimental treatment).

In developing an experiment, the theoretical construct that is the central focus of the study, as well as the measures of this CONCEPTUALIZATION, must be defined in terms of observable behaviors or events. This process

is termed *operationalization*. Then, participants are assigned randomly to conditions of the experiment, which are defined by different levels of the IV. Sometimes, more than one IV is employed; typically in such cases, each level of one IV is combined with each level of the other(s), resulting in a FACTORIAL DESIGN. Ideally, equal numbers of participants are assigned to each condition formed by the factorial combination. After assignment, participants are exposed to the particular treatment (IV) or combination of treatments (if multiple IVs are used), and the effects of this exposure are observed on the DV.

The laboratory experiment is a highly constrained research technique. Through randomization, participants theoretically are equivalent in all respects, except that some have been exposed to the IV and others have not (or they were systematically exposed

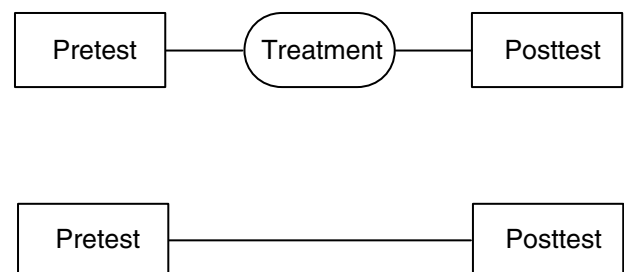


Figure 1 Example of a Pretest/Posttest/Control Group Experimental Design

NOTE: Removing the pretest from treatment and control groups results in a posttest-only/control group design. Random assignment of participants to conditions is assumed in either case.

to different values or levels of the IV). As such, they should not differ on the DV *unless* the IV has had an effect. The high degree of constraint characteristic of laboratory experimental research has caused some to question the generalizability of findings generated in these contexts. Conversely, the randomization and constraint that characterize the laboratory experiment help shield the research from extraneous variables that may affect the DV but that have not been accounted for. The trade-off of generalizability versus control helps define the most appropriate occasion for laboratory experimentation.

The laboratory experiment is the ideal method to determine cause-effect relationships. It is most profitably employed when considerable knowledge of the phenomenon under study already exists. It is not a particularly efficient method to develop hypotheses. However, if the theory surrounding a phenomenon is relatively well developed, and fine-grained hypotheses can be developed on the basis of theory, the laboratory experiment can materially assist understanding the precise nature of the construct in question. It is an ideal method to test alternative explanations of specific predicted causal relationship, facilitating identification of factors that amplify or impede a given outcome.

—William D. Crano

REFERENCE

Crano, W. D., & Brewer, M. B. (2002). *Principles and methods of social research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

LAG STRUCTURE

TIME-SERIES models typically assume that the independent variables affect the dependent variable both instantaneously and over time. This relationship is modeled with a lag structure whose parameters are then estimated by some appropriate method. Focusing on a single independent variable, x , and omitting the error term, the most general lag structure is

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} \dots$$

It is impossible to estimate the parameters of such a model because there are an infinite number of them.

A finite distributed lag structure assumes that after some given lag—say, K — x no longer affects y . Here,

the lag structure has only a finite number of parameters, that is, $\beta_{K+1} = \beta_{K+2} = \dots = 0$. Although this can be estimated in principle, MULTICOLLINEARITY makes estimation of each parameter very difficult unless it is assumed that only one or two lags of x affect current y (and even here multicollinearity may still be a problem). Thus, analysts usually impose some parametric structure on the β s. The most common assumption is that they follow a polynomial structure, such as one that is quadratic in time, as in $\beta_k = \lambda_0 + \lambda_1 k + \lambda_2 k^2$. This restriction means that analysts need only estimate three parameters.

Many analysts prefer to allow for an infinite (or at least unlimited) number of lags of x to affect y , feeling that it is odd to allow for impacts from up to some pre-specified lag and then zero impacts thereafter. The most common parameterization for this infinite distributed lag is the geometric lag structure, where the impact of x on y is assumed to decline geometrically, that is, $\beta_k = \beta \rho^k$. This can be algebraically transformed into a model with a lagged dependent variable. The assumption that the impact of x on y declines geometrically with lag length is often plausible. If it is too constraining, analysts can combine the finite and infinite lag structures by allowing the (very) few first lags to be freely estimated, as in the finite lag structure model, with the remaining lags assumed to decline geometrically. This may provide a nice balance between the flexibility of the finite lag model and the simplicity of the infinite geometric lag model without forcing the analyst to estimate too many parameters or falling afoul of multicollinearity problems.

With several independent variables, analysts must choose between the simplicity of forcing them all to follow the same lag structure and the flexibility of allowing for each variable to have its own lag structure. The price of the latter flexibility is that analysts have to estimate many more parameters. Analysts can also use lag structures to model the error process. All the various tools available to time-series analysts, such as correlograms, can be usefully combined with theoretical insight to allow analysts to choose a good lag structure.

—Nathaniel Beck

REFERENCES

Beck, N. (1991). Comparing dynamic specifications: The case of presidential approval. *Political Analysis*, 3, 51–87.

Harvey, A. (1990). *The econometric analysis of time series* (2nd ed.). Cambridge: MIT Press.
 Mills, T. C. (1990). *Time series techniques for economists*. Cambridge, UK: Cambridge University Press.

LAMBDA

The lambda coefficient is a measure of strength of relationship between NOMINAL variables. Also known as Guttman's coefficient of predictability, lambda measures the extent to which prediction of a DEPENDENT VARIABLE can be improved by knowledge of the category of a nominal PREDICTOR VARIABLE. It is a proportionate reduction in error measure because it shows how much knowing the value of one variable decreases the prediction error for the other variable. Lambda is an asymmetrical coefficient because its value depends on which of the two variables is the dependent variable because it will generally have a different value when one variable is the dependent variable and the other variable is the independent variable. Reversing the independent and dependent variables changes the value of lambda:

$$\lambda = (\sum f_i - M)/(N - M),$$

where f_i is the largest cell frequency within category i of the independent variable, M is the largest marginal frequency on the dependent variable, and N is the total number of cases. The denominator is the number of errors in predicting the dependent variable that would be made without knowing the category of the independent variable for that case, whereas the numerator shows how many fewer errors in predicting the dependent variable would be made when knowing the category of the independent variable.

Lambda varies in value between 0 and 1 and does not take on negative values. Whereas most measures of relationship view statistical INDEPENDENCE as their definition of a zero relationship between two variables, lambda instead finds no relationship whenever the modal response on the dependent variable is the same for every category of the independent variable. For example, in analyzing the relationship between the region of the country in which a person lives and the person's favorite color, the lambda value would be 0 if, in every region, more people choose blue as their favorite color than any other color.

Table 1 Favorite Color by Region, No Gain in Prediction

	<i>Northeast</i>	<i>South</i>	<i>West</i>
Blue	60	40	80
Red	20	30	10
Gray	20	30	10

NOTE: $\lambda = 0$ (predicting color from region).

Table 2 Favorite Color by Region, Complete Gain in Prediction

	<i>Northeast</i>	<i>South</i>	<i>West</i>
Blue	100	0	0
Red	0	0	100
Gray	0	100	0

NOTE: $\lambda = 1$ (predicting color from region).

Lambda also has the value of 0 whenever the two variables are statistically independent.

Lambda attains a value of 1 under "conditional unanimity." That is, it interprets a perfect relationship between two variables as the situation in which, for each independent variable category, all cases fall in the same category of the dependent variable. If everyone in the Northeast chooses blue as their favorite color, but everyone in the South chooses gray and everyone in the West chooses red, there would be perfect predictability from knowing a person's region of residence, and therefore lambda would equal 1.

There is also a SYMMETRIC version of lambda, which has the same value regardless of which variable is the dependent variable.

—Herbert F. Weisberg

REFERENCES

- Blalock, H. M. (1979). *Social statistics* (Rev. 2nd ed.). New York: McGraw-Hill.
- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association*, 49, 732–764.
- Guttman, L. (1941). An outline for the statistical theory of prediction. In P. Horst (with collaboration of P. Wallin & L. Guttman) (Ed.), *The prediction of personal adjustment* (Bulletin 48, Supplementary Study B-1, pp. 253–318). New York: Social Science Research Council.
- Weisberg, H. F. (1974). Models of statistical relationship. *American Political Science Review*, 68, 1638–1655.

LATENT BUDGET ANALYSIS

A *budget*, denoted by $\mathbf{p}$, is a vector with nonnegative components that add up to a fixed constant c . An example of a budget is the distribution of a day's time over J mutually exclusive and exhaustive activities, such as sleep, work, housekeeping, and so forth. Another example of a budget is the distribution of a \$20,000 grant over J cultural projects. Note that components of the time budget add up to 24 hours or 1,440 minutes, and the components of the financial budget add up to 20,000. Without loss of generality, the components of a budget can be divided by the fixed constant c such that they add up to 1. The time budgets of I (groups of) respondents (i.e., $\mathbf{p}_1, \dots, \mathbf{p}_I$) or the financial budgets of I \$20,000 grants can be collected in an $I \times J$ data matrix. Such data are called *compositional data*. The idea of latent budget analysis is to write the I budget as a mixture of a small number, K ($K < I$), of *latent budgets*, denoted by $\beta_1, \dots, \beta_K$.

Let Y_E be an explanatory variable (e.g., nationality, socioeconomic status, point of time) or an indicator variable with I categories, indexed by i . Let Y_R be a response variable with J categories, indexed by j , and let X be a latent variable with K categories, indexed by K . The budgets can be written as

$$\mathbf{p}_i = [P(Y_R = 1|Y_E = i), \dots, \\ P(Y_R = j|Y_E = i), \dots, \\ P(Y_R = J|Y_E = i)],$$

for $i = 1, \dots, I$, and latent budgets can be written as

$$\beta_k = [P(Y_R = 1|X = k), \dots, \\ P(Y_R = j|X = k), \dots, P(Y_R = J|X = k)],$$

for $k = 1, \dots, K$. Let $P(X = k|Y_E = i)$ denote a weight that indicates the proportion of budget $\mathbf{p}_i$ that is determined by latent budget β_k . In latent budget analysis, the budgets $\mathbf{p}_i$ ($i = 1, \dots, I$) are estimated by *expected budgets* π_i such that

$$\mathbf{p}_i \approx \pi_i = \sum_{k=1}^K P(X = k|Y_E = i) \beta_k.$$

The compositional data may be interpreted using latent budgets $\beta_1, \dots, \beta_K$ in the following procedure:

First, latent budgets $\beta_1, \dots, \beta_K$ are interpreted and entitled. Comparing the components of the budgets to the components of the independence budget $[P(Y_R = 1), \dots, P(Y_R = j)]$ can help to characterize the latent budgets. For example, if $P(Y_R = j|X = k) > P(Y_R = j)$, then β_k is characterized more than average by component j . Second, the budgets $\pi_1, \dots, \pi_I$ are interpreted on the basis of the entitled latent budgets by comparing weights $P(X = K|Y_E = i)$ to the average weight

$$P(X = k) = \sum_{i=1}^I P(X = k|Y_E = i) P(Y_E = i).$$

For example, if $P(X = k|Y_E = i) > P(X = k)$, then budget π_i is characterized more than average by latent budget β_K .

The idea of latent budget analysis was proposed by Goodman (1974), who called it asymmetric LATENT CLASS ANALYSIS; independently, the idea was also proposed by de Leeuw and van der Heijden (1988), who named it latent budget analysis. Issues with respect to latent budget analysis are discussed in van der Ark and van der Heijden (1998; graphical representation); van der Ark, van der Heijden, and Sikkel (1999; model identification); van der Heijden, Mooijaart, and de Leeuw (1992; constrained latent budget analysis); and van der Heijden, van der Ark, & Mooijaart (2002; applications).

—L. Andries van der Ark

REFERENCES

- de Leeuw, J., & van der Heijden, P. G. M. (1988). The analysis of time-budgets with a latent time-budget model. In E. Diday (Ed.), *Data analysis and informatics* (Vol. 5, pp. 159–166). Amsterdam: North-Holland.
- Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable: I. A modified latent structure approach. *American Journal of Sociology*, 79, 1179–1259.
- van der Ark, L. A., & van der Heijden, P. G. M. (1998). Graphical display of latent budget analysis and latent class analysis, with special reference to correspondence analysis. In M. Greenacre & J. Blasius (Eds.), *Visualization of categorical data* (pp. 489–508). San Diego: Academic Press.
- van der Ark, L. A., van der Heijden, P. G. M., & Sikkel, D. (1999). On the identifiability of the latent model. *Journal of Classification*, 16, 117–137.
- van der Heijden, P. G. M., Mooijaart, A., & de Leeuw, J. (1992). Constrained latent budget analysis. In P. Marsden

(Ed.), *Sociological methodology* (pp. 279–320). Cambridge, UK: Basil Blackwell.

van der Heijden, P. G. M., van der Ark, L. A., & Mooijaart, A. (2002). Some examples of latent budget analysis and its extensions. In J. A. Hagenaars & A. McCutcheon (Eds.), *Applied latent class analysis* (pp. 107–136). Cambridge, UK: Cambridge University Press.

LATENT CLASS ANALYSIS

The basic idea underlying latent class (LC) analysis is a very simple one: Some of the parameters of a postulated statistical model differ across unobserved subgroups. These subgroups form the categories of a categorical latent variable (see LATENT VARIABLE). This basic idea has several seemingly unrelated applications, the most important of which are clustering, scaling, density estimation, and random-effects modeling. Outside social sciences, LC models are often referred to as finite mixture models.

LC analysis was introduced in 1950 by Lazarsfeld, who used the technique as a tool for building typologies (or clustering) based on dichotomous observed variables. More than 20 years later, Goodman (1974) made the model applicable in practice by developing an algorithm for obtaining maximum likelihood estimates of the model parameters. He also proposed extensions for polytomous manifest variables and multiple latent variables and did important work on the issue of model identification. During the same period, Haberman (1979) showed the connection between LC models and log-linear models for frequency tables with missing (unknown) cell counts. Many important extensions of the classical LC model have been proposed since then, such as models containing (continuous) covariates, local dependencies, ordinal variables, several latent variables, and repeated measures. A general framework for categorical data analysis with discrete latent variables was proposed by Hagenaars (1990) and extended by Vermunt (1997).

Although in the social sciences, LC and finite mixture models are conceived primarily as tools for categorical data analysis, they can be useful in several other areas as well. One of these is density estimation, in which one makes use of the fact that a complicated density can be approximated as a finite mixture of simpler densities. LC analysis can also be used as a probabilistic cluster analysis tool for continuous observed

variables, an approach that offers many advantages over traditional cluster techniques such as K-means clustering (see LATENT PROFILE MODEL). Another application area is dealing with unobserved heterogeneity, for example, in regression analysis with dependent observations (see NONPARAMETRIC RANDOM-EFFECTS MODEL).

THE CLASSICAL LC MODEL FOR CATEGORICAL INDICATORS

Let X represent the latent variable and Y_ℓ one of the L observed or manifest variables, where $1 \leq \ell \leq L$. Moreover, let C be the number of latent classes and D_ℓ the number of levels of Y_ℓ . A particular LC is enumerated by the index x , $x = 1, 2, \dots, C$, and a particular value of Y_ℓ by y_ℓ , $y_\ell = 1, 2, \dots, D_\ell$. The vector notation $\mathbf{Y}$ and $\mathbf{y}$ is used to refer to a complete response pattern.

To make things more concrete, let us consider the following small data set obtained from the 1987 General Social Survey:

Y_1	Y_2	Y_3	Frequency	$P(X = 1 \mathbf{Y} = \mathbf{y})$	$P(X = 2 \mathbf{Y} = \mathbf{y})$
1	1	1	696	.998	.002
1	1	2	68	.929	.071
1	2	1	275	.876	.124
1	2	2	130	.168	.832
2	1	1	34	.848	.152
2	1	2	19	.138	.862
2	2	1	125	.080	.920
2	2	2	366	.002	.998

The three dichotomous indicators Y_1 , Y_2 , and Y_3 are the responses to the statements “allow anti-religionists to speak” (1 = allowed, 2 = not allowed), “allow anti-religionists to teach” (1 = allowed, 2 = not allowed), and “remove anti-religious books from the library” (1 = do not remove, 2 = remove). By means of LC analysis, it is possible to identify subgroups with different degrees of tolerance toward anti-religionists.

The basic idea underlying any type of LC model is that the probability of obtaining response pattern $\mathbf{y}$, $P(\mathbf{Y} = \mathbf{y})$, is a weighted average of the C class-specific probabilities $P(\mathbf{Y} = \mathbf{y} | X = x)$; that is,

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^C P(X = x) P(\mathbf{Y} = \mathbf{y} | X = x). \quad (1)$$

Here, $P(X = x)$ denotes the proportion of persons belonging to LC x .

In the classical LC model, this basic idea is combined with the assumption of LOCAL INDEPENDENCE. The L manifest variables are assumed to be mutually independent within each LC, which can be formulated as follows:

$$P(\mathbf{Y} = \mathbf{y} | X = x) = \prod_{\ell=1}^L P(Y_{\ell} = y_{\ell} | X = x). \quad (2)$$

After estimating the conditional response probabilities $P(Y_{\ell} = y_{\ell} | X = x)$, comparing these probabilities between classes shows how the classes differ from each other, which can be used to name the classes. Combining the two basic equations (1) and (2) yields the following model for $P(\mathbf{Y} = \mathbf{y})$:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^C P(X = x) \prod_{\ell=1}^L P(Y_{\ell} = y_{\ell} | X = x).$$

A two-class model estimated with the small example data set yielded the following results:

	$X = 1$ (Tolerant)	$X = 2$ (Intolerant)
$P(X = x)$	.62	.38
$P(Y_1 = 1 X = x)$	.96	.23
$P(Y_2 = 1 X = x)$	.74	.04
$P(Y_3 = 1 X = x)$	.92	.24

The two classes contain 62% and 38% of the individuals, respectively. The first class can be named “Tolerant” because people belonging to that class have much higher probabilities of selecting the tolerant responses on the indicators than people belonging to the second “Intolerant” class.

Similarly to CLUSTER ANALYSIS, one of the purposes of LC analysis might be to assign individuals to latent classes. The probability of belonging to LC x —often referred to as posterior membership probability—can be obtained by the Bayes rule:

$$\begin{aligned} P(X = x | \mathbf{Y} = \mathbf{y}) & \quad (3) \\ &= \frac{P(X = x) P(\mathbf{Y} = \mathbf{y} | X = x)}{P(\mathbf{Y} = \mathbf{y})}. \end{aligned}$$

The most common classification rule is modal assignment, which amounts to assigning each individual to the LC with the highest $P(X = x | \mathbf{Y} = \mathbf{y})$. The class membership probabilities reported in the first table show that people with at least two tolerant responses are classified into the “Tolerant” class.

LOG-LINEAR FORMULATION OF THE LC MODEL

Haberman (1979) showed that the LC model can also be specified as a LOG-LINEAR MODEL for a table with missing cell entries or, more precisely, as a model for the expanded table, including the latent variable X as an additional dimension. The relevant log-linear model for $P(X = x, \mathbf{Y} = \mathbf{y})$ has the following form:

$$\begin{aligned} \ln P(X = x, \mathbf{Y} = \mathbf{y}) &= \beta + \beta_x^X \\ &+ \sum_{\ell=1}^L \beta_{y_{\ell}}^{Y_{\ell}} + \sum_{\ell=1}^L \beta_{x, y_{\ell}}^{X, Y_{\ell}}. \end{aligned}$$

It contains a main effect, the one-variable terms for the latent variable and the indicators, and the two-variable terms involving X and each of the indicators. Note that the terms involving two or more manifest variables are omitted because of the local independence assumption.

The connection between the log-linear parameters and the conditional response probabilities is as follows:

$$P(Y_{\ell} = y_{\ell} | X = x) = \frac{\exp(\beta_{y_{\ell}}^{Y_{\ell}} + \beta_{x, y_{\ell}}^{X, Y_{\ell}})}{\sum_{r=1}^{D_{\ell}} \exp(\beta_r^{Y_{\ell}} + \beta_{x, r}^{X, Y_{\ell}})}.$$

This shows that the log-linear formulation amounts to specifying a logit model for each of the conditional response probabilities.

The type of LC formulation that is used becomes important if one wishes to impose restrictions. Although constraints on probabilities can sometimes be transformed into constraints on log-linear parameters and vice versa, there are many situations in which this is not possible.

MAXIMUM LIKELIHOOD ESTIMATION

Let I denote the total number of cells entries (or possible answer patterns) in the L -way frequency table, so that $I = \prod_{\ell=1}^L D_{\ell}$, and let i denote a particular cell entry, n_i the observed frequency in cell i , and $P(\mathbf{Y} = \mathbf{y}_i)$ the probability of having the response pattern of cell i .

The parameters of LC models are typically estimated by means of MAXIMUM LIKELIHOOD (ML). The kernel of the log-likelihood function that is maximized equals

$$\ln L = \sum_{i=1}^I n_i \ln P(\mathbf{Y} = \mathbf{y}_i).$$

Notice that only nonzero observed cell entries contribute to the log-likelihood function, a feature that

is exploited by several more efficient LC software packages that have been developed within the past few years.

One of the problems in the estimation of LC models is that model parameters may be nonidentified, even if the number of DEGREES OF FREEDOM is larger or equal to zero. Nonidentification means that different sets of parameter values yield the same maximum of the log-likelihood function or, worded differently, that there is no unique set of parameter estimates. The formal identification check is via the information matrix, which should be positive definite. Another option is to estimate the model of interest with different sets of starting values. Except for local solutions (see below), an identified model gives the same final estimates for each set of the starting values.

Although there are no general rules with respect to the identification of LC models, it is possible to provide certain minimal requirements and point to possible pitfalls. For an unrestricted LC analysis, one needs at least three indicators, but if these are dichotomous, no more than two latent classes can be identified. One has to watch out with four dichotomous variables, in which case the unrestricted three-class model is not identified, even though it has a positive number of degrees of freedom. With five dichotomous indicators, however, even a five-class model is identified. Usually, it is possible to achieve identification by constraining certain model parameters; for example, the restrictions $P(Y_\ell = 1|X = 1) = P(Y_\ell = 2|X = 2)$ can be used to identify a two-class model with two dichotomous indicators.

A second problem associated with the estimation of LC models is the presence of local maxima. The log-likelihood function of a LC model is not always concave, which means that hill-climbing algorithms may converge to a different maximum depending on the starting values. Usually, we are looking for the global maximum. The best way to proceed is, therefore, to estimate the model with different sets of random starting values. Typically, several sets converge to the same highest log-likelihood value, which can then be assumed to be the ML solution. Some software packages have automated the use of several sets of random starting values to reduce the probability of getting a local solution.

Another problem in LC modeling is the occurrence of boundary solutions, which are probabilities equal to 0 (or 1) or log-linear parameters equal to minus (or plus) infinity. These may cause numerical

problems in the estimation ALGORITHMS, occurrence of local solutions, and complications in the computation of standard errors and number of degrees of freedom of the goodness-of-fit tests. Boundary solutions can be prevented by imposing constraints or by taking into account other kinds of prior information on the model parameters.

The most popular methods for solving the ML estimation problem are the expectation-maximization (EM) and Newton-Raphson (NR) algorithms. EM is a very stable iterative method for ML estimation with incomplete data. NR is a faster procedure that, however, needs good starting values to converge. The latter method makes use the matrix of second-order derivatives of the log-likelihood function, which is also needed for obtaining standard errors of the model parameters.

MODEL SELECTION ISSUES

The goodness of-fit of an estimated LC model is usually tested by either the Pearson or the likelihood ratio chi-squared statistic (see CATEGORICAL DATA ANALYSIS). The latter is defined as

$$L^2 = 2 \sum_{i=1}^I n_i \ln \frac{n_i}{N \cdot P(\mathbf{Y} = \mathbf{y}_i)},$$

where N denotes the total sample size. As in log-linear analysis, the number of degrees of freedom (df) equals the number of cells in the frequency table minus 1, $\prod_{\ell=1}^L D_\ell - 1$, minus the number of independent parameters. In an unrestricted LC model,

$$df = \prod_{\ell=1}^L D_\ell - C \cdot \left[1 + \sum_{\ell=1}^L (D_\ell - 1) \right].$$

Although it is no problem to estimate LC models with 10, 20, or 50 indicators, in such cases, the frequency table may become very sparse and, as a result, asymptotic p values can longer be trusted. An elegant but somewhat time-consuming solution to this problem is to estimate the p values by parametric bootstrapping. Another option is to assess model fit in lower order marginal tables (e.g., in the two-way marginal tables).

It is not valid to compare models with C and $C + 1$ classes by subtracting their L^2 and df values because this conditional test does not have an asymptotic chi-squared distribution. This means that alternative methods are required for comparing models with different numbers of classes. One popular method is

the use of information criteria such as the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). Another more descriptive method is a measure for the proportion of total association accounted for by a C class model, $[L^2(1) - L^2(C)]/L^2(1)$, where the L^2 value of the one-class (independence) model, $L^2(1)$, is used as a measure of total association in the L -way frequency table.

Usually, we are not only interested in goodness of fit but also in the performance of the modal classification rule (see equation (3)). The estimated proportion of classification errors under modal classification equals

$$E = \sum_{i=1}^I \frac{n_i}{N} \{1 - \max[P(X = x | Y = y_i)]\}.$$

This number can be compared to the proportion of classification errors based on the unconditional probabilities $P(X = x)$, yielding a reduction of errors measure λ :

$$\lambda = 1 - \frac{E}{\max[P(X = x)]}.$$

The closer this nominal R^2 -type measure is to 1, the better the classification performance of a model.

EXTENSIONS OF THE LC MODEL FOR CATEGORICAL INDICATORS

Several extensions have been proposed of the basic LC model. One of the most important extensions is the inclusion of covariates or grouping variables that describe (or predict) the latent variable X . This is achieved by specifying a multinomial logit model for the probability of belonging to LC x ; that is,

$$P(X = x | Z = z) = \frac{\exp(\gamma_x^X + \sum_{k=1}^K \gamma_x^{X, Z_k} \cdot z_k)}{\sum_{r=1}^C \exp(\gamma_r^X + \sum_{k=1}^K \gamma_r^{X, Z_k} \cdot z_k)},$$

where z_k denotes a value of covariate k .

Another important extension is related to the use of information on the ordering of categories. Within the log-linear LC framework, ordinal constraints can be imposed via ASSOCIATION MODEL structures for the two variable terms $\beta_{x, y_\ell}^{X, Y_\ell}$. For example, if Y_ℓ is an ordinal indicator, we can restrict $\beta_{x, y_\ell}^{X, Y_\ell} = \beta_x^{X, Y_\ell} \cdot y_\ell$. Similar constraints can be used for the latent variable (Heinen, 1996).

In the case that a C class model does not fit the data, the local independence assumption fails to hold for one or more pairs of indicators. The common model-fitting strategy in LC analysis is to increase the number of latent classes until the local independence assumption holds. Two extensions have been developed that make it possible to follow other strategies. Rather than increasing the number of latent classes, one alternative approach is to relax the local independence assumption by including direct effects between certain indicators—a straightforward extension to the log-linear LC model. Another alternative strategy involves increasing the number of latent variables instead of the number of latent classes. This so-called LC factor analysis approach (Magidson & Vermunt, 2001) is especially useful if the indicators measure several dimensions.

Other important extensions involve the analysis of longitudinal data (see LATENT MARKOV MODEL) and partially observed indicators. The most general model that contains all models discussed thus far as special cases is the structural equation model for categorical data proposed by Hagenaars (1990) and Vermunt (1997).

OTHER TYPES OF LC MODELS

Thus far, we have focused on LC models for categorical indicators. However, the basic idea of LC analysis, that parameters of a statistical model differ across unobserved subgroups, can also be applied with variables of other scale types. In particular, three important types of applications of LC or finite mixture models fall outside the categorical data analysis framework: clustering with continuous variables, density estimation, and random-effects modeling.

Over the past 10 years, there has been a renewed interest in LC analysis as a tool for cluster analysis with continuous indicators. The LC model can be seen as a probabilistic or model-based variant of traditional nonhierarchical cluster analysis procedures such as the K-means method. It has been shown that such a LC-based clustering procedure outperforms the more ad hoc traditional methods. The method is known under names such as LATENT PROFILE MODEL, *mixture-model clustering*, *model-based clustering*, *latent discriminant analysis*, and *LC clustering*. The basic formula of this model is similar to one given in equation (1); that is,

$$f(Y = y) = \sum_{x=1}^C P(X = x) f(Y = y | X = x).$$

As shown by this slightly more general formulation, the probabilities $P(\dots)$ are replaced by densities $f(\dots)$. With continuous variables, the class-specific densities $f(\mathbf{Y} = \mathbf{y} | X = x)$ will usually be assumed to be (restricted) multivariate normal, where each LC has its own mean vector and covariance matrix. Note that this is a special case of the more general principle of density estimation by finite mixtures of simple densities.

Another important application of LC analysis is as a NONPARAMETRIC RANDOM-EFFECTS MODEL. The idea underlying this application is that the parameters of the regression model of interest may differ across unobserved subgroups. For this kind of analysis, often referred to as LC regression analysis, the LC variable serves the role of a moderating variable. The method is very similar to regression models for repeated measures or two-level data sets, with the difference that no assumptions are made about the distribution of the random coefficients.

SOFTWARE

The first LC program, MLLSA, made available by Clifford Clogg in 1977, was limited to a relatively small number of nominal variables. Today's programs can handle many more variables, as well as other scale types. For example, the LEM program (Vermunt, 1997) provides a command language that can be used to specify a large variety of models for categorical data, including LC models. Mplus is a command language-based structural equation modeling package that implements some kinds of LC models but not for nominal indicators. In contrast to these command language programs, Latent GOLD is a program with an SPSS-like user interface that is especially developed for LC analysis. It implements the most important types of LC models, deals with variables of different scale types, and extends the basic model to include covariates, local dependencies, several latent variables, and partially observed indicators.

—Jeroen K. Vermunt and Jay Magidson

REFERENCES

- Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable: I. A modified latent structure approach. *American Journal of Sociology*, 79, 1179–1259.
- Haberman, S. J. (1979). *Analysis of qualitative data: Vol. 2. New developments*. New York: Academic Press.
- Hagenaars, J. A. (1990). *Categorical longitudinal data: Loglinear analysis of panel, trend and cohort data*. Newbury Park, CA: Sage.
- Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.
- Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks, CA: Sage.
- Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis & the interpretation and mathematical foundation of latent structure analysis. In S. A. Stouffer (Ed.), *Measurement and prediction* (pp. 362–472). Princeton, NJ: Princeton University Press.
- Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology*, 31, 223–264.
- Vermunt, J. K. (1997). *Log-linear models for event histories*. Thousand Oaks, CA: Sage.

LATENT MARKOV MODEL

The latent Markov model (LMM) can either be seen as an extension of the LATENT CLASS model for the analysis of longitudinal data or as an extension of the discrete time MARKOV CHAIN model for dealing with measurement error in the observed variable of interest. It was introduced in 1955 by Wiggins and also referred to as latent transition or the hidden Markov model. The LMM can be used to separate true systematic change from spurious change resulting from measurement error and other types of randomness in the behavior of individuals.

Suppose a single categorical variable of interest is measured at T occasions, and Y_t denotes the response at occasion t , $1 \leq t \leq T$. This could, for example, be a respondent's party preference measured at 6-month intervals. Let D denote the number of levels of each Y_t , and let y_t denote a particular level, $1 \leq y_t \leq D$. Let X_t denote an occasion-specific latent variable, C a number of categories of X_t , and x_t a particular LC class at occasion t , $1 \leq x_t \leq C$. The corresponding LMM has the form

$$P(\mathbf{Y} = \mathbf{y}) = \sum_x P(X_1 = x_1) \times \prod_{t=2}^T P(X_t = x_t | X_{t-1} = x_{t-1}) \times \prod_{t=1}^T P(Y_t = y_t | X_t = x_t).$$

The two basic assumptions of this model are that the transition structure for the latent variables has the form of a first-order Markov chain and that each occasion-specific observed variable depends only on the corresponding latent variable. For identification and simplicity of the results, it is typically assumed that the error component is time homogeneous:

$$P(Y_t = y_t | X_t = x_t) = P(Y_{t-1} = y_t | X_{t-1} = x_t),$$

for $2 \leq t \leq T$. If no further constraints are imposed, one needs at least three time points to identify the LMM. Typical constraints on the latent transition probabilities are time homogeneity and zero restrictions.

The enormous effect of measurement error in the study of change can be illustrated with a hypothetical example with $T = 3$ and $C = D = 2$. To illustrate this, suppose $P(X_1 = 1) = .80$, $P(X_t = 2 | X_{t-1} = 1) = P(X_t = 2 | X_{t-1} = 2) = .10$, and $P(Y_t = 1 | X_t = 1) = P(Y_t = 2 | X_t = 2) = .20$. If we estimate a stationary manifest first-order Markov model for the hypothetical table, we find .68 in the first state at the first time point and transition probabilities of .29 and .48 out of the two states. These are typical biases encountered when not taking measurement error into account: The size of the smaller group is overestimated, the amount of change is overestimated, and there seems to be more change in the small than in the large group.

It is straightforward to extend the above single-indicator LMM to multiple indicators. Another natural extension is the introduction of covariates or grouping variables explaining individual differences in the initial state and the transition probabilities. The independent classification error (ICE) assumption can be relaxed by including direct effects between indicators at different occasions. Furthermore, mixed variants of the LMM have been proposed, such as models with MOVER-STAYER structures. In social sciences, the LMM is conceived as a tool for categorical data analysis. However, as in standard LATENT CLASS ANALYSIS, these models can be extended easily to other scale types.

The PANMARK program can be used to estimate the more standard LMMs. The LEM program can also deal with more extended models, such as models containing covariates and direct effects between indicators.

—Jeroen K. Vermunt

(Eds.), *Analyzing social and political change: A casebook of methods* (pp. 171–197). London: Sage.

MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov models and other types of models for discrete-valued time series*. London: Chapman & Hall.

Vermunt, J. K. (1997). *Log-linear models for event histories*. Thousand Oaks, CA: Sage.

Wiggins, L. M. (1973). *Panel analysis*. Amsterdam: Elsevier.

LATENT PROFILE MODEL

The latent profile model is a LATENT VARIABLE model with a categorical latent variable and continuous manifest indicators. It was introduced in 1968 by Lazarsfeld and Henry. Although under different names, very similar models were proposed in the same period by Day (1969) and Wolfe (1970).

Over the past 10 years, there has been renewed interest in this type of latent variable model, especially as a tool for CLUSTER ANALYSIS. The latent profile model can be seen as a probabilistic or model-based variant of traditional nonhierarchical cluster analysis procedures such as the K-means method. It has been shown that such a model-based clustering procedure outperforms the more ad hoc traditional methods. It should be noted that only a few authors use the term *latent profile model*. More common names are *mixture of normal components*, *mixture model clustering*, *model-based clustering*, *latent discriminant analysis*, and *latent class clustering*.

Possible social science applications of the latent profile model include building typologies and constructing diagnostic instruments. A sociologist might use it to build a typology of countries based on a set of socioeconomic and political indicators. A psychologist may apply the method to combine various test scores into a single diagnostic instrument.

As in LATENT CLASS ANALYSIS, latent profile analysis assumes that the population consists of C unobserved subgroups that can be referred to as latent profiles, latent classes, or mixture components. Because the indicators are continuous variables, it is most natural to assume that their conditional distribution is normal. The most general model is obtained with unrestricted multivariate normal distributions; that is,

$$f(\mathbf{y}) = \sum_{x=1}^C P(x) f(\mathbf{y} | \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x).$$

REFERENCES

- Langeheine, R., & Van de Pol, F. (1994). Discrete-time mixed Markov latent class models. In A. Dale & R. B. Davies

The equation states that the joint density of the L indicators, $f(\mathbf{y})$, is a mixture of class-specific densities. Each latent class x has its own mean vector μ_x and covariance matrix Σ_x . The proportion of persons in each of the components is denoted by $P(x)$. It should be noted that the model structures resembles quadratic discriminant analysis, with the important difference, of course, that the classes (groups) are unknown.

Several special cases are obtained by restricting the covariance matrix Σ_x . Common restrictions are equal covariance matrices across classes, diagonal covariance matrices, and both equal and diagonal covariance matrices. The assumption of equal covariance matrices is similar to the basic assumption of linear discriminant analysis. Specifying the covariance matrices to be diagonal amounts to assuming LOCAL INDEPENDENCE. Such a model can also be formulated as follows:

$$f(\mathbf{y}) = \sum_{x=1}^C P(x) \prod_{\ell=1}^L f(y_{\ell} | \mu_{\ell x}, \sigma_{\ell x}^2).$$

Assuming local independence and equal error variances, $\sigma_{\ell x}^2 = \sigma_{\ell}^2$, yields a specification similar to K-means clustering. In most applications, however, this local independence specification is much too restrictive.

Recently, various methods have been proposed for structuring the class-specific covariance matrices. A relatively simple one, which is a compromise between a full and a diagonal covariance matrix, is the use of block-diagonal matrices. This amounts to relaxing the local independence assumption for subsets of indicators. More sophisticated methods make use of principal component and factor analysis decompositions of the covariance matrix, as well as structural equation model-type restrictions on the Σ_x .

Another recent development is a model for combinations of continuous and categorical indicators. This is actually a combination of the classical LATENT CLASS model with the latent profile model. Another recent extension is the inclusion of covariates to predict class membership.

Software packages that can be used for estimating latent profile models are EMMIX, Mclust, Mplus, and Latent GOLD.

—Jeroen K. Vermunt

REFERENCES

Day, N. E. (1969). Estimating the components of a mixture of two normal distributions. *Biometrika*, 56, 463–474.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and application to clustering*. New York: Marcel Dekker.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge, UK: Cambridge University Press.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixtures. *Multivariate Behavioral Research*, 5, 329–350.

LATENT TRAIT MODELS

A latent trait model is a measurement model that describes the relationship between the probability of a particular score on an item (e.g., correct/incorrect, rating on an ordered scale) and the LATENT VARIABLE or the latent trait that explains this score. For an individual responding to a particular item, this probability depends on his or her latent trait value and properties of the item, such as its difficulty and its discrimination power. The term *latent trait model*, a synonym of *item response model*, is not in as much use today.

—Klaas Sijtsma

See also ITEM RESPONSE THEORY

LATENT VARIABLE

Many constructs that are of interest to social scientists cannot be observed directly. Examples are preferences, attitudes, behavioral intentions, and personality traits. Such constructs can only be measured indirectly by means of observable indicators, such as questionnaire items designed to elicit responses related to an attitude or preference. Various types of scaling techniques have been developed for deriving information on unobservable constructs of interest from the indicators. An important family of scaling methods is formed by latent variable models.

A latent variable model is a possibly nonlinear PATH ANALYSIS or graphical model. In addition to the manifest variables, the model includes one or more unobserved or latent variables representing the constructs of interest. Two assumptions define the causal mechanisms underlying the responses. First, it is assumed that the responses on the indicators are

the result of an individual's position on the latent variable(s). The second assumption is that the manifest variables have nothing in common after controlling for the latent variable(s), which is often referred to as the axiom of LOCAL INDEPENDENCE.

The two remaining assumptions concern the distributions of the latent and manifest variables. Depending on these assumptions, one obtains different kinds of latent variable models. According to Bartholomew and Knott (1999), the four main kinds are as follows: FACTOR ANALYSIS (FA), latent trait analysis (LTA), latent profile analysis (LPA), and LATENT CLASS ANALYSIS (LCA) (see table).

Manifest Variables	Latent Variables	
	Continuous	Categorical
Continuous	Factor analysis	Latent profile analysis
Categorical	Latent trait analysis	Latent class analysis

In FA and LTA, the latent variables are treated as continuous normally distributed variables. In LPA and LCA, on the other hand, the latent variable is discrete and therefore assumed to come from a multinomial distribution. The manifest variables in FA and LPA are continuous. In most cases, their conditional distribution given the latent variables is assumed to be normal. In LTA and LCA, the indicators are dichotomous, ordinal, or nominal categorical variables, and their conditional distributions are assumed to be binomial or multinomial.

The more fundamental distinction in Bartholomew and Knott's (1999) typology is the one between continuous and discrete latent variables. A researcher has to decide whether it is more natural to treat the underlying latent variable(s) as continuous or discrete. However, as shown by Heinen (1996), the distribution of a continuous latent variable model can be approximated by a discrete distribution. This shows that the distinction between continuous and discrete latent variables is less fundamental than one might initially think.

The distinction between models for continuous and discrete indicators turns out not to be fundamental at all. The specification of the conditional distributions of the indicators follows naturally from their scale types. The most recent development in latent variable modeling is to allow for a different distributional form for each indicator. These can, for example, be normal, student, log-normal, gamma, or exponential

distributions for continuous variables; binomial for dichotomous variables; multinomial for ordinal and nominal variables; and Poisson, binomial, or negative-binomial for counts. Depending on whether the latent variable is treated as continuous or discrete, one obtains a generalized form of LTA or LCA.

—Jeroen K. Vermunt and Jay Magidson

REFERENCES

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Arnold.
Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks, CA: Sage.

LATIN SQUARE

A Latin square is a square array of letters or ordered symbols in which each letter or symbol appears once and only once in each row and each column. The following are examples of 2×2 , 3×3 , and 4×4 Latin squares.

A	B	A	B	C	A	B	C	D
B	A	B	C	A	C	D	B	A
		C	A	B	D	C	A	B

The name *Latin square* was given to the squares by the famous Swiss mathematician Leonard Euler (1707–1783), who studied them and used letters of the Latin alphabet.

The three squares in the table above are called *standard squares* because their first row and first column are ordered alphabetically. Two Latin squares are *conjugate* if the rows of one square are identical to the columns of the other. For example, a 5×5 Latin square and its conjugate are as follows:

A	B	C	D	E	A	B	C	D	E
B	A	D	E	C	B	A	E	C	D
C	E	A	B	D	C	D	A	E	B
D	C	E	A	B	D	E	B	A	C
E	D	B	C	A	E	C	D	B	A

A Latin square is *self-conjugate* if the same square is obtained when its rows and columns are interchanged. The 2×2 , 3×3 , and 4×4 Latin squares above are self-conjugate.

The p different letters in a Latin square can be rearranged to produce $p!(p-1)!$ Latin squares, including the original square. For example, there are $3!(3-1)! = 12$ three-by-three Latin squares. An enumeration of Latin squares of size 2×2 through 7×7 has been made by Fisher and Yates (1934), Norton (1939), and Sade (1951).

Latin squares are interesting mathematical objects and also useful in designing experiments. A *Latin square design* enables a researcher to test the hypothesis that p population means are equal while isolating the effects of two nuisance variables, with each having p levels. Nuisance variables are undesired sources of variation in an experiment. The effects of the two nuisance variables are isolated by assigning one nuisance variable to the rows of a Latin square and the other nuisance variable to the columns of the square. The p levels of the treatment are assigned to the Latin letters of the square. The isolation of two nuisance variables often reduces the MEAN SQUARE ERROR and results in increased power.

Consider the following example. A researcher is interested in comparing the effectiveness of three diets in helping obese teenage boys lose weight. The independent variable is the three kinds of diets; the dependent variable is the amount of weight loss 3 months after going on a diet. Instead of denoting the diets by the letters A , B , and C , we will follow contemporary usage and denote the diets by the lowercase letter a and a number subscript: a_1 , a_2 , and a_3 . The researcher believes that ease in losing weight is affected by the amount that a boy is overweight and his genetic predisposition to be overweight. A rough measure of the latter variable can be obtained by asking a boy's parents if they were overweight as teenagers: c_1 denotes neither parent overweight, c_2 denotes one parent overweight, and c_3 denotes both parents overweight. This nuisance variable can be assigned to the columns of a 3×3 Latin square. The other nuisance variable, the amount that boys are overweight, can be assigned to the rows of the Latin square: b_1 is 20 to 39 pounds, b_2 is 40 to 59 pounds, and b_3 is 60 or more pounds. A diagram of this Latin square design is as follows:

	c_1	c_2	c_3
b_1	a_1	a_2	a_3
b_2	a_2	a_3	a_1
b_3	a_3	a_1	a_2

By isolating the effects of two nuisance variables, the researcher can obtain a more powerful test of the independent variable. An important assumption of Latin square designs is that there are no INTERACTIONS among the nuisance and independent variables. If the assumption is not satisfied, statistical tests will be BIASED.

François Cretté de Palluel (1741–1798), an agronomist, is credited with performing the first experiment in 1788 that was based on a Latin square. His experiment involved four breeds of sheep that were fed four diets and slaughtered on the 20th of 4 consecutive months. The person most responsible for promoting the use of experimental designs based on a Latin square was Ronald A. Fisher (1890–1962), an eminent British statistician who worked at the Rothamsted Experimental Station north of London. Fisher showed that a Latin square or combination of squares could be used to construct a variety of complex experimental designs. For a description of these designs, see Kirk (1995, chaps. 8, 13, and 14).

—Roger E. Kirk

REFERENCES

- Fisher, R. A., & Yates, F. (1934). The six by six Latin squares. *Proceedings of the Cambridge Philosophical Society*, 30, 492–507.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
- Norton, H. W. (1939). The 7×7 squares. *Annals of Eugenics*, 9, 269–307.
- Sade, A. (1951). An omission in Norton's list of 7×7 squares. *Annals of Mathematical Statistics*, 22, 306–307.

LAW OF AVERAGES

RANDOM ERROR decreases proportionally as the number of trials increases. So, by the law of averages, after enough trials, the expected outcome is more likely to be achieved. Take the example of tossing a coin. Out of 100 tosses, one expects 50 heads, but, by chance, it will probably be off that number. Say it was off by 5, by chance. If the number of tosses went up to 1,000, the expectation would again be for half heads. But again, by chance, it would probably be off by some number, say 8. The absolute number of errors has increased, from 5 to 8, but the percentage of tosses

in error has decreased, from 5% to eight tenths of 1%. In general, the law of averages says that as the number of tosses increases, the overall percentage of random error decreases. Thus, the distribution of the coin toss gets closer to its expected value of 50% heads.

—Michael S. Lewis-Beck

LAW OF LARGE NUMBERS

One of the most important results in mathematical statistics describing the behavior of observations of a random variable (i.e., observations from surveys based on random SAMPLES) is the law of large numbers. The law of large numbers concerns the difference between the EXPECTED VALUE of a variable and its AVERAGE computed from a large sample. The law states that this difference is unlikely to be large.

The simplest form of the law applies to an event (such as observing that a respondent has a certain characteristic) A , the PROBABILITY of the event (which can be identified with the relative size of the fraction of the population possessing this characteristic) $P(A)$, and its RELATIVE FREQUENCY after n observations, $r(n, A)$. The law of large numbers says that for large enough n , $r(n, A)$ will get arbitrarily close to $P(A)$, with an arbitrarily large probability. More precisely, for arbitrary (small) values a and b , there exists a threshold such that for all sample sizes n greater than this threshold,

$$P(|P(A) - r(n, A)| < a) > 1 - b.$$

The absolute deviation between the probability and the relative frequency of the event is smaller than a with a probability larger than $1 - b$, if the sample size n is large enough. There, however, remains a small probability (less than b) that the deviation is more than a .

This result is important in two aspects. First, the frequentist approach speaks of probability only if the relative frequencies show a kind of stability (i.e., the relative frequencies computed for large sample sizes tend to be close to each other) and postulates that the relative frequencies are manifestations of the directly unobservable probability. The fact that this property is obtained as a result shows that the mathematical theory of probability appropriately incorporates the fundamental assumptions. Second, the law of large numbers

is also a fundamental result of mathematical statistics. In this context, it says that with large enough sample sizes, the estimates (relative frequency of an event) are very likely to be close to the true value (probability of the event).

The previous result does not indicate how large a sample size is needed to get close to the probability with the relative frequency. The following version is more helpful in this aspect. Let $E(X)$ and $D(X)$ be the expected value and the STANDARD DEVIATION of a variable. Then, for arbitrary positive value c ,

$$P(|X - E(X)| > cD(X)) < 1/c^2.$$

This result measures the deviation between the observed value (e.g., the relative frequency) and the expected value (e.g., the probability) in units of the standard deviation. The probability that the deviation exceeds c times the standard deviation is less than $1/c^2$. This result applies to any quantity that may be observed in a survey.

For example, a poll is used in a town to estimate the fraction of those who support a certain initiative. If the sample size is, say, 100, the following bound is obtained for the likely behavior of the estimate. From the BINOMIAL distribution, the standard deviation is not more than 5%, and the probability that the result of the poll deviates from the truth by more than 10% (i.e., 2 standard deviations) is less than $1/4 = 0.25$.

—Tamás Rudas

REFERENCE

Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed.). New York: John Wiley.

LAWS IN SOCIAL SCIENCE

The question of the possibility of social laws has divided social scientists along both disciplinary and epistemological lines. Macroeconomics has traditionally depended on laws, or at least lawlike statements (e.g., the “laws of supply and demand”), whereas most historians would be skeptical of the existence of historical laws. These positions (to which of course there are exceptions) reflect the epistemological foundations of the disciplines.

The epistemological grounds for the rejection of laws can be summarized as the view that there is

too much variability deriving from individual consciousness and free will in the social world to permit lawlike regularities. Although the conclusion of this argument may be ultimately correct, it is also often premised on an oversimplified version of what a “law” is. The common notion of a scientific law is one that brooks no exceptions in the predicted behavior of phenomena. Classical laws of physics are often cited, such as gravity, thermodynamics, and so forth. However, most scientific laws have quite different characteristics, and even classical laws will exhibit certain local exceptions and, under certain circumstances (outside of human experience), may break down altogether (Nagel, 1979, chaps. 4, 5).

Laws can express fundamental theoretical principles (such as those in economics), they may be mathematical axioms or experimentally established invariances, or they may express statistical regularities. Although a theory is usually seen to be a more speculative statement about regularities, sometimes the difference between a law and a theory may be no more than a linguistic or social convenience. Thus, although they both enjoy similar degrees of corroboration, one speaks of Newton’s laws and Einstein’s theories (Williams, 2000, p. 33). If, then, by accepting the possibility of theoretical corroboration, one accepts there can be regularities in the social world, does this entail a commitment to laws that all have the property of claiming the existence of some nonaccidental occurrences? Perhaps regularities in economic and demographic behavior are as predictable, under given circumstances, as those of gases. Statistical laws, at least, would seem to be possible. However, to continue the gas analogy, although certain local conditions (such as temperature or pressure variations) will create statistical fluctuation in the distribution of gases, the characteristics of the exceptions to demographic or economic laws are much more dependent on unique cultural or historical circumstances (Kincaid, 1994). Thus, there will be regularity in the exceptions to physical laws, but demonstrating such regularity in social laws requires reference to other universalist principles such as rationality.

What is perhaps more important than whether there can be social laws is that good evidence suggests that there can be nonaccidental regularities in the social world, even though such regularities may vary in their cultural or historical applicability and extent.

—Malcolm Williams

REFERENCES

- Kincaid, H. (1994). Defending laws in the social sciences. In M. Martin & L. McIntyre (Eds.), *Readings in the philosophy of social science* (pp. 111–130). Cambridge: MIT Press.
- Nagel, E. (1979). *The structure of science: Problems in the logic of scientific explanation*. Indianapolis, IN: Hackett.
- Williams, M. (2000). *Science and social science: An introduction*. London: Routledge Kegan Paul.

LEAST SQUARES

The least squares method—a very popular technique—is used to compute estimations of parameters and to fit data. It is one of the oldest techniques of modern statistics, being first published in 1805 by the French mathematician Legendre in a now-classic memoir. But this method is even older because it turned out that, after the publication of Legendre’s memoir, Gauss, the famous German mathematician, published another memoir (in 1809) in which he mentioned that he had previously discovered this method and used it as early as 1795. A somewhat bitter anteriority dispute followed (a bit reminiscent of the Leibniz-Newton controversy about the invention of calculus), which, however, did not diminish the popularity of this technique. Galton used it (in 1886) in his work on the heritability of size, which laid down the foundations of CORRELATION and (also gave the name) REGRESSION analysis. Both Pearson and Fisher, who did so much in the early development of statistics, used and developed it in different contexts (FACTOR ANALYSIS for Pearson and EXPERIMENTAL DESIGN for Fisher).

Nowadays, the least squares method is widely used to find or estimate the numerical values of the parameters to fit a function to a set of data and to characterize the statistical properties of estimates. It exists with several variations: Its simpler version is called ORDINARY LEAST SQUARES (OLS); a more sophisticated version is called WEIGHTED LEAST SQUARES (WLS), which often performs better than OLS because it can modulate the importance of each observation in the final solution. Recent variations of the least squares method are alternating least squares (ALS) and PARTIAL LEAST SQUARES (PLS).

FUNCTIONAL FIT EXAMPLE: REGRESSION

The oldest (and still most frequent) use of OLS was LINEAR REGRESSION, which corresponds to the problem

of finding a line (or curve) that best fits a set of data. In the standard formulation, a set of N pairs of observations $\{Y_i, X_i\}$ is used to find a function giving the value of the dependent variable (Y) from the values of an independent variable (X). With one variable and a linear function, the prediction is given by the following equation:

$$\hat{Y} = a + bX. \quad (1)$$

This equation involves two free parameters that specify the intercept (a) and the slope (b) of the regression line. The least squares method defines the estimate of these parameters as the values that minimize the sum of the squares (hence the name *least squares*) between the measurements and the model (i.e., the predicted values). This amounts to minimizing the following expression:

$$\varepsilon = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i [Y_i - (a + bX_i)]^2, \quad (2)$$

where ε stands for “error,” which is the quantity to be minimized. This is achieved using standard techniques from calculus—namely, the property that a quadratic (i.e., with a square) formula reaches its minimum value when its derivatives vanish. Taking the derivative of ε with respect to a and b and setting them to zero gives the following set of equations (called the *normal equations*):

$$\frac{\partial \varepsilon}{\partial a} = 2Na + 2b \sum X_i - 2 \sum Y_i = 0 \quad (3)$$

and

$$\begin{aligned} \frac{\partial \varepsilon}{\partial b} &= 2b \sum X_i^2 \\ &+ 2a \sum X_i - 2 \sum Y_i X_i = 0. \end{aligned} \quad (4)$$

Solving these two equations gives the least squares estimates of a and b as

$$a = M_Y - bM_X, \quad (5)$$

with M_Y and M_X denoting the means of Y and X , and

$$b = \frac{\sum (Y_i - M_Y)(X_i - M_X)}{\sum (X_i - M_X)^2}. \quad (6)$$

OLS can be extended to more than one independent variable (using MATRIX ALGEBRA) and to nonlinear functions.

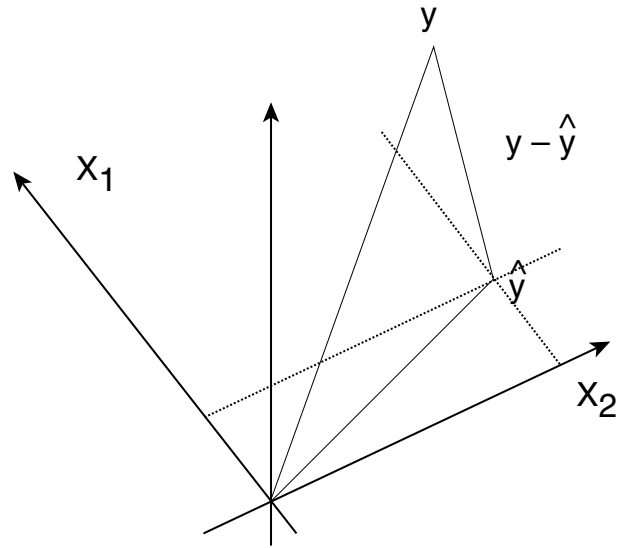


Figure 1 The Least Squares Estimate of the Data Is the Orthogonal Projection of the Data Vector Onto the Independent Variable Subspace

THE GEOMETRY OF LEAST SQUARES

OLS can be interpreted in a geometrical framework as an orthogonal projection of the data vector onto the space defined by the independent variable. The projection is orthogonal because the predicted values and the actual values are uncorrelated. This is illustrated in Figure 1, which depicts the case of two independent variables (vectors $\mathbf{x}_1$ and $\mathbf{x}_2$) and the data vector ($\mathbf{y}$) and shows that the error vector ($\mathbf{y} - \hat{\mathbf{y}}$) is orthogonal to the least squares ($\hat{\mathbf{y}}$) estimate, which lies in the subspace defined by the two independent variables.

OPTIMALITY OF LEAST SQUARES ESTIMATES

OLS estimates have some strong statistical properties. Specifically, when (a) the data obtained constitute a RANDOM SAMPLE from a well-defined POPULATION, (b) the population model is linear, (c) the error has a zero expected value, (d) the independent variables are linearly independent, and (e) the error is normally distributed and uncorrelated with the independent variables (the so-called homoskedasticity assumption), then the OLS estimate is the BEST LINEAR UNBIASED ESTIMATOR, often denoted with the acronym BLUE (the five conditions and the proof are called the Gauss-Markov conditions and theorem). In addition, when the Gauss-Markov conditions hold, OLS estimates are also MAXIMUM LIKELIHOOD estimates.

WEIGHTED LEAST SQUARES

The optimality of OLS relies heavily on the HOMOSKEDASTICITY assumption. When the data come from different subpopulations for which an independent estimate of the error variance is available, a better estimate than OLS can be obtained using weighted least squares (WLS), also called GENERALIZED LEAST SQUARES (GLS). The idea is to assign to each observation a weight that reflects the uncertainty of the measurement. In general, the weight w_i , assigned to the i th observation, will be a function of the variance of this observation, denoted σ_i^2 . A straightforward weighting schema is to define $w_i = \sigma_i^{-1}$ (but other, more sophisticated weighted schemes can also be proposed). For the linear regression example, WLS will find the values of a and b minimizing:

$$\begin{aligned}\varepsilon_w &= \sum_i w_i (Y_i - \hat{Y}_i)^2 \\ &= \sum_i w_i [Y_i - (a + bX_i)]^2.\end{aligned}\quad (7)$$

ITERATIVE METHODS: GRADIENT DESCENT

When estimating the parameters of a nonlinear function with OLS or WLS, the standard approach using derivatives is not always possible. In this case, iterative methods are very often used. These methods search in a stepwise fashion for the best values of the estimate. Often, they proceed by using at each step a linear approximation of the function and refine this approximation by successive corrections. The techniques involved are known as gradient descent and Gauss-Newton approximations. They correspond to nonlinear least squares approximation in numerical analysis and nonlinear regression in statistics. NEURAL NETWORKS constitute a popular recent application of these techniques.

—Hervé Abdi

REFERENCES

- Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: John Wiley.
- Greene, W. H. (2002). *Econometric analysis*. New York: Prentice Hall.
- Nocedal, J., & Wright, S. (1999). *Numerical optimization*. New York: Springer-Verlag.
- Plackett, R. L. (1972). The discovery of the method of least squares. *Biometrika*, 59, 239–251.
- Seal, H. L. (1967). The historical development of the Gauss linear model. *Biometrika*, 54, 1–23.

LEAST SQUARES PRINCIPLE.

See REGRESSION

LEAVING THE FIELD

Despite the enormous amount of space devoted to gaining entry in the ethnographic literature, very little attention has been paid to its corollary, *leaving the field*. Also called *disengagement*, leaving the field is the process of exiting from field settings and disengaging from the social relationships cultivated and developed during the research act.

Leaving the field is influenced by the way fieldworkers gain entry, the intensity of the social relationships they develop in the field, the research bargains they make, and the kinds of persons they become to their informants—or how the researcher and the researched become involved in what Rochford (1992) calls “a politics of experience” (p. 99).

In describing his field experiences with a Buddhist movement in America, Snow (1980) discusses three central issues: (a) information sufficiency, which refers to whether the researcher collected enough data to answer the questions posed by the research; (b) certain extraneous precipitants, which could be institutional, interpersonal, or intrapersonal; and (c) barriers that pressure the ethnographer to remain in the field, such as the feelings of the group concerning the researcher’s departure and the overall intensity of field relationships.

In some cases, leaving the field is influenced by the research bargain. This written or unwritten agreement between the researcher and a *gatekeeper* is usually more important at the beginning of the research and is largely forgotten later on. However, the nature of the bargain influences disengagement and helps to shape exit strategies.

Another related issue involves the misrepresentation of the researcher’s identity. Although most researchers are overt in their investigations, there are cases when researchers deliberately misrepresent or are unable to be truthful or are unable to maintain their commitment or promises. Informants expect the researchers to live up to their commitments permanently. When the research ends, the level of commitment dissipates, researchers return to their everyday lives, and

informants may feel betrayed and manipulated. As researchers prepare to disengage, they need to consider future relationships with their informants, especially if they want to return for further study or clarification.

Gallmeier's (1991) description of conducting FIELD RESEARCH among a group of minor-league hockey players demonstrates the advantages gained by staying in touch with informants. Such revisiting allows the ethnographer to obtain missing data as well as follow the careers of informants who have remained in the setting and thereby observe the changes that have occurred. Revisits also enable the researcher to "take the findings back into the field" (Rochford, 1992, pp. 99–100) to assess the validity of their field reports. Some ethnographers are experimenting with ways to share their findings with informants, actually soliciting their comments on the final analysis and write-up. By giving the respondents the "last word," ethnographers are employing a validity measure referred to as MEMBER VALIDATION. The strength of this approach is that it affords the researched an opportunity to respond to what has been written about them, and it assists the ethnographer in comparing the *etic* with the *emic* perspective.

—Charles P. Gallmeier

REFERENCES

- Gallmeier, C. P. (1991). Leaving, revisiting, and staying in touch: Neglected issues in field research. In W. B. Shaffir & R. A. Stebbins (Eds.), *Fieldwork experience: Qualitative approaches to social research* (pp. 224–231). London: Sage.
- Rochford, E. B., Jr. (1992). On the politics of member validation: Taking findings back to Hare Krishna. In G. Miller & J. A. Holstein (Eds.), *Perspectives on social problems* (Vol. 3, pp. 99–116). Greenwich, CT: JAI.
- Snow, D. (1980). The disengagement process: A neglected problem in participant observation research. *Qualitative Sociology*, 3, 100–122.

LESLIE'S MATRIX

If we know the size and the composition of a population as well as the age-specific fertility and survival (or, alternatively, mortality) rates at the current time, can we project the population size and composition to future times? Leslie's (1945) creative use of MATRIX ALGEBRA is intended for answering this question.

In matrix notation, the population sizes at time $t + 1$, $\mathbf{N}_{t+1}$, is the product of the Leslie matrix, $\mathbf{L}$, and the population sizes at time t , $\mathbf{N}_t$.

$$\mathbf{N}_{t+1} = \mathbf{L}\mathbf{N}_t.$$

This can be written out in vector notation as

$$\begin{bmatrix} N_{1,t+1} \\ N_{2,t+1} \\ \vdots \\ N_{J,t+1} \end{bmatrix} = \begin{bmatrix} F_1 & F_2 & F_3 & \dots & F_J \\ S_1 & 0 & 0 & \dots & 0 \\ 0 & S_2 & 0 & \dots & 0 \\ 0 & 0 & \vdots & \ddots & 0 \\ 0 & \dots & 0 & S_{j-1} & 0 \end{bmatrix} \begin{bmatrix} N_{1,t} \\ N_{2,t} \\ \vdots \\ N_{J,t} \end{bmatrix},$$

where, for $j = 1, \dots, J$, $N_{j,t}$ is the population size of the j th age group at time t , $N_{j,t+1}$ is the population size of the j th age group at time $t + 1$, S_j is the probability of an individual will survive to the next age group, and F_j is the age-specific fertility of the j th age group.

Let us use a very simple hypothetical example of a population divided into three age classes or groups, each of which has only 10 individuals—hence, $\mathbf{N}'_t = [10, 10, 10]$. Let us further assume that the three age-specific fertility rates are all 0, and the two age-specific survival probabilities are 1 (the survival probability for the last age class is, by definition, 0 as no one lives beyond the upper age limit defined by the last age class). Then, $\mathbf{N}'_{t+1} = [0, 10, 10]$ because everyone survives to the next age class except those in the oldest age class. This population will disappear from the surface of the earth by $t + 3$. Readers familiar with matrix multiplication will quickly see that the result can be obtained from the equation above.

A more realistic situation could have a 30%, 40%, and 30% fertility schedule in the first row of the Leslie matrix, as well as the survival probabilities of 1 again for the first two age groups. In this scenario, we have a constant population size and structure; that is, the three age groups will always contain 10 individuals each, maintaining a *stationary population*, a special case of the useful STABLE POPULATION MODEL. In the real world, however, populations do not follow such beautiful models, but their behaviors can nonetheless be captured by the Leslie matrix.

The Leslie matrix, although originally designed to study population changes in the human as well as the animal world, has much wider applicability in the social sciences beyond a simple projection of population sizes of a closed system. Recent applications include a sensitivity study of the U.S. population rate of growth due to immigration (Liao, 2001) and a study of the spread and prevalence of rumors known as “urban legends” (Noymer, 2001). The Leslie matrix could possibly be applied to study growth in organizations or firms where age would be replaced by seniority.

—Tim Futing Liao

REFERENCES

- Leslie, P. H. (1945). On the use of matrices in certain population mathematics. *Biometrika*, 33, 183–212.
- Liao, T. F. (2001). How responsive is U.S. population growth to immigration? A situational sensitivity analysis. *Mathematical Population Studies*, 9, 217–229.
- Noymer, A. (2001). The transmission and persistence of “urban legends”: Sociological application of age-structured epidemic models. *Journal of Mathematical Sociology*, 25, 299–323.

LEVEL OF ANALYSIS

A social science analysis can be set at the MICRO level when individuals are studied, or it can be set at the MACRO level when aggregates of individuals such as cities, counties, provinces, or nations become the UNIT OF ANALYSIS. There is no cut-and-dried definition, and a researcher may choose to analyze a level of analysis between the very micro and the very macro, such as households or neighborhoods.

—Tim Futing Liao

LEVEL OF MEASUREMENT

The level of measurement defines the relation among the values assigned to the ATTRIBUTES of a VARIABLE. Social science researchers conventionally use four such levels: nominal, ordinal, interval, and ratio. For defining these four types of measurement, an increasing number of ASSUMPTIONS are required. For instance, a NOMINAL VARIABLE needs only discrete categories for representing its attributes, an ORDINAL MEASURE has attributes that are ordered, an INTERVAL

measure requires equal distancing among them, and a RATIO SCALE also has a meaningful absolute zero.

—Tim Futing Liao

LEVEL OF SIGNIFICANCE

The level of significance refers to a particular value or level of PROBABILITY a researcher sets in statistical SIGNIFICANCE TESTING of a HYPOTHESIS, commonly known as ALPHA SIGNIFICANCE LEVEL OF A TEST. By social science convention, researchers report the level of significance at the 0.001, 0.01, 0.05, or 0.10 level of making a TYPE I ERROR, although sometimes a lower or higher level of significance is used.

—Tim Futing Liao

LEVENE'S TEST

This test, which was developed by Howard Levene (1960), is a one-way analysis of variance on the absolute deviation of each score from the mean of its group in which the scores of the groups are unrelated. It is an *F* test in which the population estimate of the variance between the groups is compared with the population estimate of the variance within the groups.

$$F = \frac{\text{Between-groups variance estimate}}{\text{Within-groups variance estimate}}.$$

The larger the between-groups variance estimate is in relation to the within-groups variance estimate, the larger the *F* value is and the more likely it is to be statistically significant. A statistically significant *F* value means that the variances of the groups are not equal or homogeneous. These variances may be made to be more equal by transforming the original scores such as taking their square root or logarithm.

Analysis of variance is usually used to determine whether the means of one or more factors and their interactions differ significantly from that expected by chance. One of its assumptions is that the variances of the groups should be equal so that they can be pooled into a single estimate of the variance within the groups. However, if one or more of the groups have a larger variance than the other groups, the larger variance will increase the within-groups variance, making it less likely that an effect will be statistically significant.

Several tests have been developed to determine whether group variances are homogeneous. Some of

these tests require the number of cases in each group to be the same or are sensitive to data that are not normally distributed. Levene's test was designed to be used with data that are not normally distributed and where group size is unequal. Milliken and Johnson (1984, p. 22) recommend Levene's test when the data are not normally distributed. Levene's test is the only homogeneity of variance test currently offered in the statistical software of SPSS for Windows.

In factorial analysis of variance, in which there is more than one unrelated factor, Levene's test treats each cell as a group in a one-way analysis of variance. For example, in a factorial analysis of variance consisting of one factor with two groups or levels and another factor with three groups or levels, there will be six ($2 \times 3 = 6$) cells. These six cells will be analyzed as a one-way analysis of variance for unrelated scores. In analysis of covariance, the absolute deviation is that between the original score and the score predicted by the unrelated factor and the covariate.

Rosenthal and Rosnow (1991, p. 340) have recommended that Levene's test can also be used to test the homogeneity of related scores when a single-factor repeated-measures analysis of variance is carried out on the absolute deviation of each score from its group mean.

—Duncan Cramer

REFERENCES

- Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278–292). Stanford, CA: Stanford University Press.
- Milliken, G. A., & Johnson, D. A. (1984). *Analysis of messy data: Vol. 1. Designed experiments*. New York: Van Nostrand Reinhold.
- Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research* (2nd ed.). New York: McGraw-Hill.

LIFE COURSE RESEARCH

A key concern of many social scientists is to understand major life events—which can be as simple as completing education, marriage, first birth, and employment—and, in particular, the transition into a particular event as well as the relations between these events throughout one's life span or life course. Life course researchers typically take a dynamic view, as any life course event can be the consequence of past experience and future expectation. Life course research

uses many types of methodology, both qualitative and quantitative, such as the LIFE HISTORY METHOD and EVENT HISTORY ANALYSIS.

—Tim Futing Liao

REFERENCE

- Giele, J. Z., & Elder, G. H. (1998). *Methods of life course research: Qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

LIFE HISTORY INTERVIEW. See LIFE STORY INTERVIEW

LIFE HISTORY METHOD

A life story is the story a person tells about the life he or she has lived. For Atkinson (1998), it is “a fairly complete narrating of one's entire experience of life as a whole, highlighting the most important aspects” (p. 8). In social science, life stories exist in many forms: long and short, specific and general, fuzzy and focused, surface and deep, realist and romantic, modernist and postmodernist. And they are denoted by a plethora of terms: *life stories*, *life histories*, *life narratives*, *self stories*, *mystories*, *autobiographies*, *auto/biographies*, *oral histories*, *personal testaments*, and *life documents*. All have a slightly different emphasis and meaning, but all focus on the first persona accounting of a life. (For a discussion of this diversity, see Denzin, 1989, pp. 27–49.)

An initial, simple, but helpful contrast can be made between “long” and “short” life stories. *Long life stories* are perhaps seen as the key to the method—the full-length book account of one person's life in his or her own words. They are usually gathered over a long period of time with gentle guidance from the researcher, often backed up with keeping diaries, intensive observation of the subject's life, interviews with friends, and perusals of letters and photographs. The first major sociological use of life histories in this way is generally credited to the publication of Thomas and Znaniecki's five-volume *The Polish Peasant in Europe and America*, which makes the famous claim that life histories “constitute the perfect type of sociological material” (Thomas & Znaniecki, 1918–1920/1958,

pp. 1832–1833). One volume of this study provides a 300-page story of a Polish émigré to Chicago, Wladek Wisniewski, written 3 months before the outbreak of World War I. In his story, Wladek describes the early phases of his life in the Polish village of Lubotyn, where he was born the son of a rural blacksmith. He talks of his early schooling, his entry to the baker's trade, his migration to Germany to seek work, and his ultimate arrival in Chicago and his plight there. Twenty years on this "classic" study was honored by the Social Science Research Council "as the finest exhibit of advanced sociological research and theoretical analysis" (see Blumer, 1939/1969, p. vi). Following from this classic work, life histories became an important tool in the work of both Chicago and Polish sociologists. By contrast, *short life stories* take much less time, tend to be more focused, and are usually published as one in a series. They are gathered through in-depth interviews, along with open-ended questionnaires, requiring gentle probes that take somewhere between half an hour and 3 hours. The stories here usually have to be more focused than the long life histories.

COMPREHENSIVE, TOPICAL, AND EDITED

Making another set of distinctions, Allport (1942) suggested three main forms of life history writing: the comprehensive, the topical, and the edited. The *comprehensive life document* aims to grasp the totality of a person's life—the full sweep—from birth to his or her current moment, capturing the development of a unique human being. The *topical life document* does not aim to grasp the fullness of a person's life but confronts a particular issue. The study of Stanley, The Jack Roller, in the late 1920s focuses throughout on the delinquency of his life (Shaw, 1966). The document is used to throw light on a particular topic or issue. The *edited life document* does not leave the story quite so clearly in the subject's voice. Rather, the author speaks and edits the subjects into his or her account—these documents are often used more for illustration. The classic version of this is William James's (1902/1952) *The Varieties of Religious Experience: A Study in Human Nature*. In this book-length study, organized around themes such as "Conversion," "Saintliness," and "Mysticism," James draws from series of case studies of religious experience. No one life dominates—instead, extracts appear in many places.

NATURALISTIC, RESEARCHED, AND REFLEXIVE

A further distinction that is increasingly made lies in what can be called *naturalistic*, *researched*, and *reflexive* life stories, although there can be overlaps between them. *Naturalistic life stories* are stories that naturally occur, which people just tell as part of their everyday lives, unshaped by the social scientist. Naturalistic life stories are not artificially assembled but just happen in situ. They are omnipresent in those most popular forms of publishing: confessions, personal testaments, and autobiographies. Classically, these may include Rousseau's (1781/1953) *Confessions* or slave narratives such as Harriet A. Jacobs' *Incidents in the Life of a Slave Girl: Written by Herself* (1861). They are first-person accounts unshaped by social scientists.

By contrast, *researched life stories* are specifically gathered by researchers with a wider—usually social science—goal in mind. These do not naturalistically occur in everyday life; rather, they have to be interrogated out of subjects. ORAL HISTORY, sociological LIFE HISTORY, literary biographies, psychological case studies—all these can bring life stories into being that would not otherwise have happened. Both the case studies of Stanley and Wladek cited above would be classic instances of this. The role of researchers is crucial to this activity: Without them, there would be no life story. In their questioning, they shape and generate the stories.

Reflexive and recursive life stories differ from the other two types of life stories by bringing with them a much greater awareness of their own construction and writing. Most earlier life stories would convey a sense that they were telling the story of a life, but these latter kinds of stories become self-conscious and suggest that storytelling is a language game, an act of speech, a mode of writing, a social construction. Whatever else they may say, stories do not simply tell the tale of a life. Here, a life is seen more richly as "composed" or constructed: The writer becomes a part of the writing. In this view, life story research usually involves a storyteller (the narrator of the story) and an interviewer (who acts as a guide in this process), as well as a narrative text that is assembled: producer, teller, and text. The producer and the teller together construct the story in specific social circumstances. Stories can be told in different ways at different times. Stanley (1992) even finds that so many of the most common distinctions that are made—for

example, between biography and autobiography—are false: “Telling apart fiction, biography and autobiography is no easy matter” (p. 125). Because of this, she carves out yet another word, *auto/biography*—“a term which refuses any easy distinction between biography and autobiography, instead recognizing their symbiosis”—and applies it to some of her own life (Stanley, 1992, p. 127).

Although life stories have often been criticized for being overly subjective, too idiographic, and nongeneralizable, it is noticeable that in the closing years of the 20th century, such research was becoming more and more popular.

—Ken Plummer

See also AUTOBIOGRAPHY, AUTOETHNOGRAPHY, CASE STUDY, HUMANISM AND HUMANISTIC RESEARCH, INTERPRETATIVE BIOGRAPHY, INTERVIEWING IN QUALITATIVE RESEARCH, LIFE COURSE RESEARCH, LIFE HISTORY INTERVIEW, NARRATIVE ANALYSIS, NARRATIVE INTERVIEW, ORAL HISTORY, QUALITATIVE RESEARCH, REFLEXIVITY

REFERENCES

- Atkinson, R. (1998). *The life story interview* (Qualitative Research Methods Series, Vol. 44). Thousand Oaks, CA: Sage.
- Blumer, H. (1969). *An empirical appraisal of Thomas and Znaniecki (1918–20) The Polish peasant in Europe and America*. New Brunswick, NJ: Transaction Books. (Original work published 1939)
- Denzin, N. K. (1989). *Interpretative biography*. London: Sage.
- Jacobs, H. A. (1861). *Incidents in the life of a slave girl: Written by herself*. Boston.
- James, W. (1952). *The varieties of religious experience: A study in human nature*. London: Longmans, Green & Co. (Original work published 1902)
- Plummer, K. (2001). *Documents of life 2: An invitation to a critical humanism*. London: Sage.
- Roberts, B. (2002). *Biographical research*. Buckingham, UK: Open University Press.
- Rousseau, J.-J. (1953) *The confessions*. Harmondsworth, UK: Penguin. (Original work published 1781.)
- Shaw, C. (1966). *The jack roller*. Chicago: University of Chicago Press.
- Smith, S., & Watson, J. (2001). *Reading autobiography: A guide for interpreting life narratives*. Minneapolis: University of Minnesota Press.
- Stanley, L. (1992). *The auto/biographical I: The theory and practice of feminist auto/biography*. Manchester, UK: Manchester University Press.
- Thomas, W. I., & Znaniecki, F. (1918–1920). *The Polish peasant in Europe and America* (5 vols.). New York: Dover.
- Thomas, W. I., & Znaniecki, F. (1958). *The Polish peasant in Europe and America* (2 vols.). New York: Dover. (Original five-volume work published in 1918–1920)

LIFE STORY INTERVIEW

A life story is the story a person chooses to tell about the life he or she has lived as completely and honestly as possible, what that person remembers of it, and what he or she wants others to know of it, usually as a result of a guided INTERVIEW by another; it is a fairly complete narrating of one's entire experience of life as a whole, highlighting the most important aspects. A life story is the narrative essence of what has happened to a person. It can cover the time from birth to the present, or before and beyond. It includes the important events, experiences, and feelings of a lifetime.

Life story, LIFE HISTORY, and ORAL HISTORY can be different terms for the same thing. The difference between them is often emphasis and scope. Although a life history or an oral history often focuses on a specific aspect of a person's life, such as work life or a special role played in some part of the life of a community, a life story interview has as its primary and sole focus a person's entire life experience. A life story brings order and meaning to the life being told, for both the teller and the listener. It is a way to better understand the past and the present, as well as a way to leave a personal legacy for the future. A life story gives us the vantage point of seeing how one person experiences and subjectively understands his or her own life over time. It enables us to see and identify the threads that connect one part of one's life to another, from childhood to adulthood.

HISTORICAL AND DISCIPLINARY CONTEXT

Stories, in traditional communities of the past, played a central role in the lives of the people. The stories we tell of our lives today are guided by the same vast catalog of enduring elements, such as archetypes and motifs, that are common to all human beings. These expressions of ageless experiences take a range of forms in different times and settings and come together to create a pattern of life, with many repetitions, that becomes the basis for the plot of the story of a life. Our life stories follow such a pattern or blueprint and represent, each in its own way, a balance between

the opposing forces of life. It is within this ageless and universal context that we can best begin to understand the importance and power of the life story interview and how it is fundamental to our very nature. Our life stories, just as myths did in an earlier time, can serve the classic functions of bringing us into accord with ourselves, others, the mystery of life, and the universe around us.

Researchers in many academic disciplines have been interviewing others for their life stories, or for some aspect of their lives, for longer than we recognize. The life story interview has evolved from the oral history, life history, and other ETHNOGRAPHIC and field approaches. As a QUALITATIVE RESEARCH method for looking at life as a whole and as a way of carrying out an in-depth study of individual lives, the life story interview stands alone. It has a range of interdisciplinary applications for understanding single lives in detail and how the individual plays various roles in society.

Social scientists—such as Henry A. Murray, who studied individual lives using life narratives primarily to understand personality development, and Gordon Allport, who used personal documents to study personality development in individuals—began using a life-as-a-whole approach in the 1930s and 1940s. This approach reached its maturation in studies of Luther and Gandhi by Erik Erikson (1975), who used the life history to explore how the historical moment influenced lives.

Today, social scientists reflect a broad interest in narrative as it serves to illuminate the lives of persons in society. The narrative study of lives is a significant interdisciplinary methodological approach (represented by the series edited by Josselson & Lieblich, 1993–1999) that aims to further the theoretical understanding of individual life narratives through in-depth studies, methodological examinations, and theoretical explorations.

As a research tool that is gaining much interest and use in many disciplines today, there are two primary approaches to life stories: CONSTRUCTIONIST and NATURALISTIC. Some narrative researchers conceive of the life story as a circumstantially mediated, constructive collaboration between the interviewer and interviewee. This approach stresses the situated emergence of the life story as opposed to the subjectively faithful, experientially oriented account. In the constructionist perspective, life stories are evaluated not so much for how well they accord with the life experiences in question but more in terms of how accounts of lives are used

by a variety of others, in addition to the subjects whose lives are under consideration, for various descriptive purposes (Holstein & Gubrium, 2000).

The approach used at the University of Southern Maine Center for the Study of Lives is more naturalistic and has evolved from an interdisciplinary context, taking in folklore, counseling psychology, and cross-cultural human development and exploring how cultural values and traditions influence development across the life cycle from the subjective representation of one's life. This approach seeks to understand another's lived experience, as well as his or her relation to others, from that person's own voice. If we want to know the unique perspective of an individual, there is no better way to get this than in his or her voice. This allows that person, as well as others, to see his or her life as a whole, across time—and the parts of it may all fit together as one continuous whole, or some parts may seem discontinuous with others, or both might be possible, as well.

Life stories have gained respect and acceptance in many academic circles. Psychologists see the value of personal narratives in understanding development and personality. Anthropologists use the life history, or individual CASE STUDY, as the preferred unit of study for their measures of cultural similarities and variations. Sociologists use life stories to understand and define relationships as well as group interactions and memberships. In education, life stories have been used as a new way of knowing and teaching. Literary scholars use autobiography as texts through which to explore questions of design, style, content, literary themes, and personal truth. In using the oral history approach, historians find that life story materials are an important source for enhancing local history.

USES OF THE LIFE STORY INTERVIEW

The life story is inherently interdisciplinary; its many research uses can help the teller as well as the scholar understand a very broad range of social science issues better. A life story narrative can be a valuable text for learning about the human endeavor, or how the self evolves over time, and becomes a meaning maker with a place in society, the culture, and history. A life story can be one of the most emphatic ways to answer the question, "Who am I?"

Life stories can help the researcher become more aware of the range of possible roles and standards that exist within a human community. They can define

an individual's place in the social order of things and can explain or confirm experience through the moral, ethical, or social context of a given situation. They can provide the researcher with information about a social reality existing outside the story, described by the story.

Life stories can provide clues to what people's greatest struggles and triumphs are, where their deepest values lie, what their quest has been, where they might have been broken, and where they were made whole again. Life stories portray religion, spirituality, worldview, beliefs, and community making as lived experience.

The research applications of the life story interview are limitless. In any field, the life story itself could serve as the centerpiece for published research, or segments could be used as data to illustrate any number of research needs. The life story interview allows for more data than the researcher may actually use, which not only is good practice but also provides a broad foundation of information to draw on. The life story approach can be used within many disciplines, as well as for many substantive issues. To balance out the databases that have been relied on for so long in generating theory, researchers need to record more life stories of women and members of diverse groups. The feminine voice needs to be given more opportunities to be heard and understood. Because how we tell our stories is mediated by our culture, we need to hear the life stories of individuals from underrepresented groups, to help establish a balance in the literature and expand the awareness and knowledge level of us all.

THE METHODOLOGY OF THE LIFE STORY INTERVIEW

Although a fairly uniform research methodology can be applied, and much important data can be gotten from a life story, there is a level of subjectivity, even chance, involved in doing a life story interview. The same researcher may use different questions with different interviewees, based on a number of variables, and still end up with a fairly complete life story of the person being interviewed. Different interviewers may also use different questions, depending on the particular focus of their project. The life story interview is essentially a template that will be applied differently in different situations, circumstances, or settings.

For example, more than 200 questions suggested in *The Life Story Interview* (Atkinson, 1998) can be asked to obtain a life story. These questions are not meant to

be used in their entirety or as a structure that is set in stone. They are suggested questions only, with only the most appropriate few to be used for each person interviewed. There are times when a handful of the questions might actually be used and other times when two or three dozen might be applied; in each case, a different set of questions most likely would be chosen. The best interview is usually the result of carefully chosen OPEN-ENDED QUESTIONS that elicit and encourage detailed, in-depth responses.

The life story interview can be *approached* scientifically, but it is best *carried out* as an art. Although there may be a structure (a set of questions, or parts thereof) that can be used, each interviewer will apply this in his or her own way. The execution of the interview, whether structured or not, will vary from one interviewer to another. The particular interviewee is another important factor. Some life storytellers offer highly personal meanings, memories, and interpretations of their own, adding to the artful contours of their life story, but others may need more help and guidance filling the details.

A life story interview, as carried out at the Center for the Study of Lives, involves the following three steps: (a) *planning* (preinterview)—preparing for the interview, including understanding why a life story can be beneficial; (b) *doing the interview itself* (interviewing)—guiding a person through the telling of his or her life story while recording it on either audiotape or videotape; and (c) *transcribing and interpreting the interview* (postinterview)—leaving out questions and comments by the interviewer, as well as other repetitions (only the words of the person telling his or her story remain, so that it then becomes a flowing, connected narrative in that person's own words). One might then give the transcribed life story to the person to review and check over for any changes he or she might want to make in it. What we will end up with is a flowing life story in the words of the person telling it. The only editing necessary would be to delete repetitions or other completely extraneous information.

The length of a life story interview can vary considerably. Typically, for the kind of life story interviewing being described here, at least two or three interviews with the person are needed, with each one lasting 1 to 1½ hours. This length of interview can usually provide more than enough information to gain a good understanding of the person's life or the research topic.

One characteristic of the life story interview is that it is not a standardized research instrument. Although

framed as both an art and a science and requiring much preinterview planning and preparation, the life story interview is a highly contextualized, personalized approach to gathering qualitative information about the human experience. It demands many spontaneous, individual judgments on the part of the interviewer while the interview is in progress. Its direction can be determined in the spur of the moment by unexpected responses to questions or by the way a life is given its particular narrative structure. The quest in a life story interview is to present the unique voice and experience of the storyteller, which may also merge at some points with the universal human experience.

ETHICAL AND INTERPRETATION ISSUES

A number of important conceptual and interpretive issues also need to be considered. Perhaps most important, because we are asking real people to tell us their true stories, and we are attempting to assist and collaborate with them in this process and then take their story to a larger audience, we have to be able to address satisfactorily the important ethical questions about maintaining a consistency between our original intention and the final product. The ethics of doing a life story interview are all about being fair, honest, clear, and straightforward. A relationship develops that is founded on a moral responsibility to ensure that the interests, rights, and sensitivities of the storyteller are protected.

Also important are the interpretive challenges of doing a life story interview. The ultimate aim of the narrative investigation of a human life, using the life stories approach, is the interpretation of experience, a complex matter because both *interpretation* and *experience* are highly relative terms. Subjectivity is thus at the center of the process of life storytelling. Categories of analysis will emerge from a review of each life story text itself, along with a complexity of patterns and meanings, rather than being set from the beginning, as in quantitative studies. A fundamental interpretive guideline is that the storyteller should be considered both the expert and the authority on his or her own life. This demands a standard of RELIABILITY and VALIDITY that is appropriate to the life story interview as a subjective reflection of the experience in question.

There are a multiplicity of perspectives possible, and the narrative arrived at by different interviewers will be representative of a position, just as a portrait painted from the side or from the front is still a faithful

portrait. Historical truth may not be the main issue in personal narrative; what matters is if the life story is deemed trustworthy, more than true.

Nevertheless, internal consistency is still important. What is said in one part of the narrative should not contradict what is said in another part. There *are* inconsistencies in life, and people may react to this one way one time and a different way at another time, but their stories of what happened and what they did should be consistent within themselves.

Whether it is for research purposes on a particular topic or question or to learn more about human lives and society from one person's perspective, life stories serve as excellent means for understanding how people see their own experiences, their own lives, and their interactions with others. The approach of the life story interview suggested here may avoid many of the typical research and publication dilemmas if certain primary values are kept in mind. Setting out to help people tell their stories in their own words gives the researcher a very clear intent, as well as final product.

More life stories need to be brought forth that respect and honor the personal meanings the life storytellers give to their stories. Interpretation and analysis of, as well as commentary on, the life story, even though highly individualized and very subjective, can follow the same approach whether this is carried out within the same life story or across many life stories. In most cases and disciplines, the way to best understand a life story is to find the meaning that comes with it, emerges from it, or is waiting to be drawn out from it.

Life storytelling is a process of personal meaning making. Often, the telling of the story can bring about the knowing of it for the first time, and the text of the story can carry its own meaning. Beyond that, each discipline or research project may have its own theoretical or substantive interest being looked for in the story, and this is when the analysis of a life story or a series of life stories can be examined from theoretical or thematic perspectives. The real value of a life story interview is that it allows for and sheds light on the common life themes, values, issues, and struggles that many people might share, as well as on the differences that do exist between people.

—Robert Atkinson

REFERENCES

- Atkinson, R. (1998). *The life story interview* (Qualitative Research Methods Series, Vol. 44). Thousand Oaks, CA: Sage.

- Erikson, E. H. (1975). *Life history and the historical moment*. New York: Norton.
- Holstein, J. A., & Gubrium, J. F. (2000). *The self we live by: Narrative identity in a postmodern world*. New York: Oxford University Press.
- Josselson, R., & Lieblich, A. (Eds.). (1993–1999). *The narrative study of lives* (Vols. 1–6). Newbury Park, CA: Sage.
- Titon, J. (1980). The life story. *Journal of American Folklore*, 93(369), 276–292.

LIFE TABLE

Life tables are powerful and versatile analytic tools that describe the general parameters of a population's mortality regime such as its age-specific probabilities of survival and mean survival times. Although life tables are most commonly used in the analysis of mortality of human populations, they can be used for any population that is subject to duration-specific probabilities of experiencing one or more "exit events." They have been used, for example, in the study of school enrollment and graduation, labor force participation, migration, marital status, and health and disability. The main attribute of life tables is the ability to compare the parameters of mortality (or some other exit process) across populations without the need to adjust for compositional differences.

Life tables lend themselves to two different types of interpretations: the experiences of a COHORT or the experiences of a stationary population. If the data are period specific, the life table portrays the hypothetical exits of a *synthetic* cohort played out over time. If the data follow the experiences of a *true* (birth) cohort, the life table is called a "generation life table." Generation life tables of human populations are rare because they require the observation of age-specific death rates stretching over the life span of the longest lived individuals in the birth cohort.

If the parameters in a life table are interpreted as representing the mortality experiences of a stationary population, then the numbers of births or entrants into the population are assumed to remain constant with each new cohort experiencing the observed age-specific mortality or exit rates over time. Because the numbers of births, deaths, and age-specific mortality rates remain constant, the total size and the age

distribution of the hypothetical stationary population also remain constant.

THE BASIC LIFE TABLE FUNCTIONS

Life tables consist of a series of interconnected columns arrayed by age groups.

- x —exact age x . The values of x range from 0, the youngest possible age, to the oldest possible age, represented by ω . In abridged life tables, the single-year age intervals are aggregated into n -year intervals, x to $x + n$. For human populations, it is common for n to equal 1 for the first age interval, 4 for the second interval, and 5 or 10 for all but the last interval, which is open-ended.

- ${}_na_x$ —the separation fraction. It refers to the fraction of the age interval lived by persons who die between ages x and $x + n$. For all but the youngest and oldest age intervals, the values of ${}_na_x$ are often set to .50, a value that assumes deaths are spread evenly throughout the interval. Because deaths in the first years of life are concentrated near the beginning of the interval, the separation fractions for the youngest age groups are usually less than .50. There are a variety of suggested values for the separation fractions for the younger ages (see, e.g., Preston, Heuveline, & Guillot, 2001).

- ${}_nM_x$ —the observed age-specific mortality rate for the population aged x to $x + n$. In the formula below, the capital letters M , D , and N refer to the mortality rate, the number of deaths, and the size of the mid-year population between exact ages x and $x + n$, respectively. If the life table describes a national population or subpopulation, the number of deaths is usually derived from vital statistics, and the size of the population is derived from a census:

$${}_nM_x = {}_nD_x / {}_nN_x.$$

- ${}_nq_x$ —the proportion of the members of the life table cohort alive at the beginning of the age interval (at exact age x) who die before reaching the end of the age interval. It is the most crucial column in the life table. Except for the youngest and oldest age groups, values of ${}_nq_x$ are often estimated using the following formula:

$${}_nq_x = {}_nM_x / ((1/n) + ((1 - {}_na_x) \times {}_nM_x)).$$

Because no one survives past the oldest possible age (ω), ${}_nq_{\omega-n} = 1.0$. Because the

Table 1 Life Table for the U.S. Population, 2000

Age Interval	Death Rate in Population	Separation Fraction	Proportion Dying During Interval	Number Alive at Exact Age x	Number Dying During Interval	Stationary Population		
						Number Alive in Interval	Number Alive Age x and Older	Mean Number of Years of Remaining Life
x to $x + n - 1$	${}_nM_x$	${}_na_x$	${}_nq_x$	l_x	${}_nd_x$	${}_nL_x$	T_x	e_x
0–1	— ^a	0.125	0.0069	100,000	690	99,393	7,684,601	76.8
1–4	0.000328	0.35	0.0013109	99,310	130	396,902	7,585,209	76.4
5–14	0.000186	0.5	0.0018583	99,180	184	990,877	7,188,307	72.5
15–24	0.000815	0.5	0.0081169	98,996	804	985,937	6,197,430	62.6
25–34	0.001080	0.5	0.0107420	98,192	1,055	976,646	5,211,493	53.1
35–44	0.001997	0.5	0.0197726	97,137	1,921	961,769	4,234,847	43.6
45–54	0.004307	0.5	0.0421620	95,217	4,015	932,093	3,273,078	34.4
55–64	0.010054	0.5	0.0957278	91,202	8,731	868,367	2,340,986	25.7
65–74	0.024329	0.5	0.2169046	82,471	17,888	735,272	1,472,618	17.9
75–84	0.056943	0.5	0.4432345	64,583	28,625	502,703	737,346	11.4
85+	0.153244	— ^a	1.0000000	35,958	35,958	234,643	234,643	6.5

a. Value for this cell is not calculated.

under-enumeration of infants is common, ${}_1q_0$ is often derived from vital statistics data on infant deaths and births (i.e., the infant mortality rate) rather than from the age-specific death rate for infants.

- l_x —the number of persons in the life table cohort alive at exact age x . The value of l_0 is usually set to 100,000. In general, $l_{x+n} = l_x - d_x$.

- ${}_nd_x$ —the number of deaths occurring to people age x to $x + n$.

In general, ${}_nd_x = l_x \times {}_nq_x$. Because all persons alive at the beginning of the last interval die by the end of the last interval, ${}_nd_{\omega-n} = l_{\omega-n}$.

- ${}_nL_x$ —the number of person-years lived by the life table cohort between ages x and $x + n$ or the number of people alive in the stationary population between ages x and $x + n$:

$${}_nL_x = (n \cdot l_{x+n}) + ({}_na_x \cdot {}_nd_x).$$

Estimating the number of person-years lived in the last age interval (or the number of people alive in the stationary population who are age $\omega - n$ and older) can be problematic. It is customary to use the age-specific mortality rate observed in the population to estimate it:

$${}_nL_{\omega-n} = \frac{{}_nd_{\omega-n}}{{}_nM_{\omega-n}}.$$

- T_x —the total number of person-years lived by the life table cohort after age x or the total number of

people alive in the stationary population ages x and older:

$$T_x = \sum_x^{\omega-n} {}_nL_x.$$

- e_x —the average remaining lifetime for people age x :

$$e_x = \frac{\sum_x^{\omega-n} {}_nL_x}{l_x} = \frac{T_x}{l_x}.$$

The example provided in Table 1, which is based on data for the total population in the United States in 2000, is a period-specific life table. The life table parameters suggest that the members of a synthetic cohort living out their lives according to the mortality rates observed in the year 2000 will live, on average, 76.8 years, but a bit more than a third of them (35,958/100,000) will experience their 85th birthday. The stationary population generated by the life table, which will experience 100,000 births and 100,000 deaths every year, will have a constant size of 7,684,601 members, of whom 99,393 are between the exact ages of 0 and 1.

GENERAL OBSERVATIONS ABOUT LIFE TABLES

The initial size of the life table cohort is l_0 . Because all members of the life table cohort exit the life table through death, the total number of deaths ($\sum_x d_x$)

equals the initial size of the life table cohort (l_0). The total size of the stationary population is T_0 . The crude death rate (CDR) and crude birth rate (CBR) of the life table stationary population equal l_0/T_0 . The life expectancy at age 0 in the life table population (e_0) is the inverse of the crude death rate ($1/CDR$). Preston et al. (2001); Shryock, Siegel, and Associates (1973); and Smith (1992), among many others, provide detailed descriptions of life table functions for single-decrement life tables (in which there is only one exit), multiple-decrement life tables (with two or more exits), and increment and decrement life tables (which allow both entrances and exits), as well as their interpretations and uses. They also provide detailed summaries of more complex demographic methods, such as the projection and indirect estimation of populations and the use of stationary and stable population models, which build on the basic life table functions described here.

—Gillian Stevens

REFERENCES

- Preston, S. H., Heuveline, P., & Guillot, M. (2001). *Demography: Measuring and modeling population processes*. Oxford, UK: Blackwell.
- Shryock, H. S., Siegel, J. S., & Associates. (1973). *The methods and materials of demography*. New York: Academic Press.
- Smith, D. P. (1992). *Formal demography*. New York: Plenum.

LIKELIHOOD RATIO TEST

A likelihood ratio test is used to determine whether overall model fit is improved by relaxing one or more restrictions and is usually used to test whether individual coefficients or linear combinations of coefficients are equal to specific values. In general, it can be used to compare any two NESTED models (but not nonnested models). The likelihood ratio test is based on the MAXIMUM LIKELIHOOD principle of the estimates and allows a test for whether differences in the likelihood of the data for two models are systematic or due to SAMPLING ERROR. The likelihood ratio statistic is constructed using the likelihood of the unrestricted model (evaluated at the maximum likelihood parameter estimates), L^* , and the likelihood of the restricted model, LR^* , and is defined as $LR = -2 \ln(LR^*/L^*)$. This can be rewritten

as $2(\ln L^* - \ln LR^*)$. Because removing model restrictions always increases the likelihood, the likelihood ratio statistic is positive and can be shown to follow a chi-square distribution with m degrees of freedom, where m is the number of restrictions: ($LR \sim X_m^2$). Under the null hypothesis that the restricted model is just as good as the unrestricted model, the expected value of R is 0. Including considerations for sampling error, however, the expected value of LR is the number of restrictions, m . To the extent that there are systematic differences between the models, then the observed value of LR will be greater than m . Using classical hypothesis testing, significance tests can be conducted using the value of LR . ASYMPTOTICALLY, the likelihood ratio test is equivalent to both the Wald test (W) and the Lagrange multiplier test (LM), which is also known as Rao's score test. In small samples, the tests will perform differently, but in the case of linear restrictions, it can be shown that $LM < LR < W$.

—Frederick J. Boehmke

REFERENCES

- Kmenta, J. (2000). *Elements of econometrics* (2nd ed.). Ann Arbor: University of Michigan Press.
- Maddala, G. G. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.

LIKERT SCALE

The Likert (or summated rating) scale is a very popular device for measuring people's attitudes, beliefs, emotions, feelings, perceptions, personality characteristics, and other psychological constructs. It allows people to indicate their position on items along a quantitative continuum. Three features characterize this sort of SCALE. First, there are multiple items that are combined into a total score for each respondent. Second, each item has a stem that is a statement concerning some object (person, place, or thing). This might concern an aspect of the person completing the scale or an aspect of another person or object. Third, the person completing the scale is asked to choose one of several response choices that vary along a continuum. They can be unipolar, ranging from low to high, or bipolar, ranging from extreme negative to extreme positive. These responses are quantified (e.g., from

1 to 6) and summed across items, yielding an interpretable quantitative score.

The development of a Likert scale generally involves five steps (Spector, 1992). First, the construct of interest is carefully defined to clarify its nature and boundaries. Second, format decisions are made and the items are written. Third, the scale is pilot tested on a small SAMPLE to be sure the items are clear and make sense. Fourth, the scale is administered to a sample of 100 to 200 (or more) individuals. An item analysis is conducted to help choose those items that are intercorrelated with one another and form an internally consistent scale. Coefficient alpha is computed as a measure of internal consistency reliability with a target of at least .70 (Nunnally, 1978), although .80 to .90 is more desirable. Finally, research is conducted to provide data on the scale's CONSTRUCT VALIDITY to be sure it is reasonable to interpret it as reflecting the construct defined in Step 1.

The response choices in Likert scales can differ in number of type. The most popular form is the *agree-disagree*, in which respondents indicate the extent to which they agree with each stem. Usually, there are five to seven response choices—for example, *disagree (agree) very much*, *disagree (agree) somewhat*, and *disagree (agree) slightly*. Other common alternatives are frequency (e.g., *never*, *seldom*, *sometimes*, *often*) and evaluation (e.g., *poor*, *fair*, *good*, *outstanding*).

There has been some debate with bipolar scales about whether there should be an odd number of response choices with a neutral response in the middle (e.g., *neither agree nor disagree*) or an even number without the neutral response. Those advocating an odd number argue that one should not force the ambivalent person to make a choice in one direction or the other. Those advocating an even number point out that the neutral response is often misused by respondents (e.g., to indicate that the item is not applicable) and that it may encourage people to be noncommittal. There is generally little practical difference in results using even or odd numbers of response choices.

—Paul E. Spector

REFERENCES

- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Spector, P. E. (1992). *Summated rating scale construction: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-082). Newbury Park, CA: Sage.

LIMDEP

LIMDEP is an integrated computer program for the estimation and analysis of a variety of linear and nonlinear econometric and statistical models for LIMited DEpendent variables and other forms of data, such as survival and sample selection data. For further information, see the Web site of the software: <http://www.limdep.com>.

—Tim Futing Liao

LINEAR DEPENDENCY

A set of vectors is said to be linearly dependent if any one of the vectors in the set can be written as a linear combination of (some of) the others. More generally, we say that two variables are linearly dependent if

$$\lambda_1 X_1 + \lambda_2 X_2 = 0,$$

where λ_1 and λ_2 are CONSTANTS such that λ_1 or $\lambda_2 \neq 0$. This definition can be expanded to k variables such that a perfect linear relationship is said to exist when

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k = 0,$$

where $\lambda_1 \cdots \lambda_k$ are not all zero simultaneously.

To illustrate, consider the following set of data where each of the three (INDEPENDENT) VARIABLES is represented as a vector containing n (the number of observations) values:

$$\begin{bmatrix} X_1 & X_2 & X_3 \\ 4 & 16 & 12 \\ 7 & 30 & 22 \\ 12 & 34 & 29 \\ 13 & 22 & 24 \\ 1 & 50 & 26 \end{bmatrix}.$$

In this case, X_3 is a perfect linear combination of X_1 and X_2 ($X_{3i} = \frac{1}{2}X_{2i} + X_{1i}$), and therefore, a (perfect) linear dependency exists in this data set.

Linear dependency (among the independent variables) has serious implications for the estimation of regression coefficients. It is a violation of the ASSUMPTION that $\mathbf{X}$ (the MATRIX of independent variables) is a $n \times k$ matrix with rank k (i.e., the columns of $\mathbf{X}$ are linearly independent, and there are at least k observations). In addition, it leads to the commonly understood problem of (MULTI) COLLINEARITY.

The implications of linear dependency on the estimation of REGRESSION COEFFICIENTS depend on the degree of the dependency. If there is perfect linear dependency among the independent variables, the regression coefficients cannot be uniquely determined. To see this, consider the following SIMPLE regression model with two independent variables:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i.$$

Let $X_{2i} = 2X_{1i}$ (implying a perfect linear dependency between the independent variables). Then, it is the case that

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 2X_{1i} + u_i$$

or

$$y_i = \beta_0 + (\beta_1 + 2\beta_2)X_{1i} + u_i.$$

It should be clear from the last equation that in the presence of a perfect linear dependency, it is not possible to estimate the *independent* effects of X_1 (i.e., β_1) and X_2 (i.e., β_2).

A similar result is obtained in matrix form by noting that if a perfect linear dependency exists in the matrix of independent variables ($\mathbf{X}$), then both ($\mathbf{X}$) and ($\mathbf{X}'\mathbf{X}$) are singular, and neither can be inverted. Recalling that the vector of regression coefficients in matrix form is given by $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, it is clear that these coefficients cannot be estimated in the presence of perfect linear dependency.

In social science research, it is rarely the case that a perfect linear dependency between independent variables will exist. Rather, it is quite possible that something like “near-perfect” dependency will be present. Near-perfect linear dependency can result in high but not perfect (multi) collinearity.

In the presence of imperfect collinearity, regression coefficients can be estimated, and they are still BLUE (best linear unbiased estimators), as imperfect collinearity violates none of the assumptions of the classical linear regression model (CLRM). However, there remain some nontrivial consequences. Coefficients will have large VARIANCES and COVARIANCES. As a consequence, CONFIDENCE INTERVALS for the coefficients will be wider (and t -STATISTICS will often be insignificant), leading researchers to fail to reject the NULL HYPOTHESIS when it is, in fact, false (i.e., a higher incidence of TYPE II ERRORS). Finally, coefficient estimates and their STANDARD ERRORS can be sensitive to small changes in the data.

—Charles H. Ellis

REFERENCES

- Green, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River: Prentice Hall.
- Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.
- Kennedy, P. (1998). *A guide to econometrics* (4th ed.). Cambridge: MIT Press.

LINEAR REGRESSION

Linear regression refers to a linear FUNCTION expressing the RELATIONSHIP between the conditional mean of a RANDOM VARIABLE (the DEPENDENT VARIABLE) and the corresponding values of one or more explanatory variables (INDEPENDENT VARIABLES). The dependent variable is a random variable whose realization is composed of the DETERMINISTIC effects of fixed values of the explanatory variables as well as random disturbances. We could express such a relationship as $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i$.

We can speak of either a population regression function or a sample regression function. The sample regression function is typically used to estimate an unknown population regression function. If a linear regression model is well specified, then the expected value of ε_i in the population is zero. Given this expectation, UNBIASED estimates of the population regression function can be calculated from the sample by solving the problem $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$. This equation can also be thought of in terms of estimating a conditional mean as in $E(Y_i|X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$. Both equations give the average values of the STOCHASTIC dependent variable conditional on fixed values of the independent variables. The goal of linear regression analysis is to find the values $\hat{\beta}$ that best characterize $\hat{Y}_i$ or $E(Y_i|X_i)$.

The term *linear regression* implies no particular method of estimation. Rather, a linear regression can be estimated in any of several ways. The most popular approach is to use ORDINARY LEAST SQUARES (OLS). This approach does not explicitly incorporate assumptions about the probability distribution of the disturbances. However, when attempting to relate from the sample to the population regression functions, unbiased and EFFICIENT estimation requires compliance with the Gauss-Markov assumptions. Using OLS, the analyst minimizes the function defined by the sum of squared sample residuals with respect to the regression parameters. Another approach to estimating a linear

regression is to use MAXIMUM LIKELIHOOD ESTIMATION. This approach requires the assumption of normally distributed disturbances, in addition to the assumptions of OLS. Using maximum likelihood, the analyst maximizes the function defined by the product of the joint probabilities of all disturbances with respect to the regression parameters. This is called the joint likelihood function. Yet another approach is to use the method of moments. This approach uses the analogy principle whereby moment conditions are used to derive estimates.

With regard to the meaning of linearity, it is useful to clarify that we are not concerned with linearity in the variables but with linearity in the parameters. For a linear regression, we require that predicted values of the dependent variable be a linear function of the estimated parameters. For example, suppose we estimate a sample regression such as $\hat{Y}_i = \hat{\alpha} + \hat{\beta}\sqrt{X_i} + \varepsilon_i$. In this regression, the relation between $\hat{Y}_i$ and X_i is not linear, but the relation between $\hat{Y}_i$ and $\sqrt{X_i}$ is linear. If we think of $\sqrt{X_i}$ instead of X_i as our explanatory variable, then the linearity assumption is met. However, this cannot be applied equally for all estimators. For example, suppose we want to estimate a function such as

$$\hat{Y}_i = \hat{\alpha} + \frac{\sqrt{\hat{\beta}}}{\hat{\alpha}} X_i + \varepsilon_i.$$

This is not a linear regression, but it requires nonlinear regression techniques.

—B. Dan Wood and Sung Ho Park

REFERENCES

- Casella, G., & Berger, R. L. (1990). *Statistical inference*. Belmont, CT: Duxbury.
- Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.

LINEAR TRANSFORMATION

A linear transformation is a mathematical function that changes all the values a variable can take into a new set of corresponding values. When the linear transformation is imposed, the units and the shape of the variable's HISTOGRAM change. A wide range of rank-preserving transformations is used in statistics to move from one distribution shape to another. These fall into two groups: linear and nonlinear transformations, as illustrated in Table 1. The nonlinear transformations

give curvature to the relationship, whereas the linear transformations do not.

As a function, such as $X' = 4 + 0.2X$, it is important for all existing values over the range of X to map uniquely onto the new variable X' . Conversions from the Fahrenheit to Celsius temperature scales and in general from imperial to metric units qualify as linear transformations. The logarithmic transformation is a nonlinear transformation.

In general, a linear transformation of a variable X has the form

$$X' = a + bX.$$

The slope of the function with respect to X is constant. In empirical work, it is often necessary to convert earnings to a new currency—for example, $Y_{\text{euro}} = 0.66Y_{\text{£}}$ when the exchange rate is 0.66 Euros per Pound sterling, or $Y_{\text{euro}} = RY_{\text{£}}$ when the rate in general is denoted R . It is also useful to convert weekly earnings to annual earnings (e.g., $Y_{\text{yrly}} = 52 \cdot Y_{\text{wkly}}$).

More complex linear transformations use more than one source variable. Earnings that include an annual bonus may be summed up by merging the contents of the two variables:

$$Y_{\text{yrly}} = 52 \cdot (H \cdot W) + B,$$

where H = hours per week, W = wage per hour, B = bonus per year, and Y_{yrly} = total annual earnings. Y_{yrly} is a linear transformation of H if and only if W is invariant with respect to H (i.e., no overtime supplement). In general, for a variable X , the transformation function X' is a linear function if

$$\frac{D_X^2}{D_X} = 0.$$

It is possible to see scale construction as a linear transformation process with explicit weights for the different variables comprising the new scale.

In forming linear transformations, a nonzero constant can be used. A multiplicative constant other than 1 changes the VARIANCE, altering the relative dispersion of the variable vis-à-vis its own mean by a proportion. The best-known linear transformation is the standardization of a variable around its mean. To read changes in X in unit-free language (or, equivalently, in terms of standard deviation units), one subtracts X 's mean and divides X by its standard deviation

$$X'_i = \frac{(X_i - \bar{X})}{\text{sd}(X)},$$

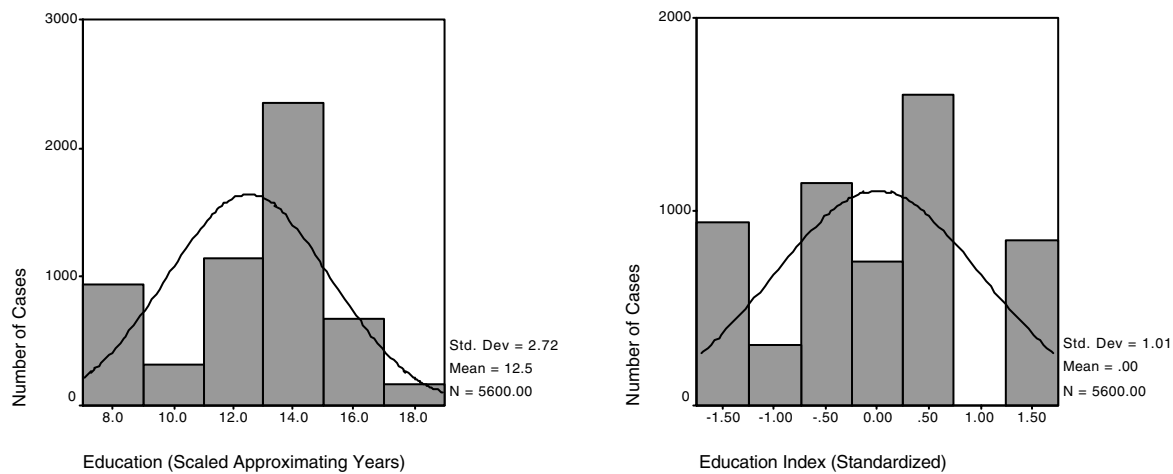


Figure 1 Distribution of a Variable and Its Standardized Variant
SOURCE: British Household Panel Survey, Wave 9, 1999–2000 (Taylor, 2000).

Table 1

	Linear Transformation	Nonlinear Transformation
Illustration using algebra	$Y' = a + bY$, where, for instance, $11 = 3 + (2 \cdot 4)$ if, in general, $Y' = 3 + 2Y$	$Y' = \exp(Y) = e^Y$ noting that $e = 2.718$, for instance, $7.4 = e^2$
Illustration using graphs		

where we define

$$sd(X) = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{n - 1}}.$$

The new variable X' is a linear transformation of X because the derivative of X' with respect to X_i is constant over the values of X . The standard deviation is a fixed parameter here. For a normally distributed variable such as levels of education (approximated in years), the standardizing transformation produces the changes illustrated in Figures 1 and 2 in the distribution of X .

The standardizing transformation is sometimes described as producing a Z-distribution if the underlying distribution of X is normal. Z-distributions are standardized normal distributions that have mean zero and a standard deviation of 1. As Figures 1 and 2 show, however, the distribution’s shape changes during this procedure.

—Wendy K. Olsen

REFERENCES

Economic and Social Research Council Research Centre on Micro-Social Change. (2001). *British Household Panel*

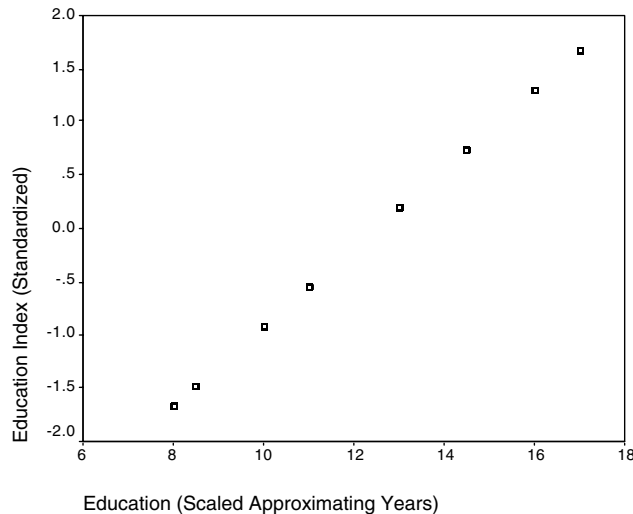


Figure 2

Survey [computer file]. Colchester, UK: The Data Archive [distributor], 28 February. Study Number 4340.

Taylor, M. F. (Ed.) (with Brice, J., Buck, N., & Prentice-Lane, E.). (2000). *British Household Panel Survey User Manual Volume B9: Codebook*. Colchester, UK: University of Essex.

LINK FUNCTION

The link function is what links the random component and the systematic component of GENERALIZED LINEAR MODELS. It is also what connects the possibly nonlinear outcome variable with the linear combination of EXPLANATORY VARIABLES.

—Tim Futing Liao

LISREL

LISREL was the first statistical software package that became widely available to social scientists for performing analysis of Linear Structural RELations models, better known as STRUCTURAL EQUATION MODELS. Today, it can analyze structural equation models of a variety of types of forms, including ordinal and longitudinal data. For further

information, see the Web site of the software: <http://www.ssicentral.com/lisrel/mainlis.htm>.

—Tim Futing Liao

LISTWISE DELETION

Listwise deletion is a term used in relation to computer software programs such as SPSS in connection with the handling of MISSING DATA. Listwise deletion of missing data means that all cases with missing data involved in an analysis will be ignored in an analysis. Consider the following scenario: We have a sample of 200 cases, we want to produce a set of PEARSON CORRELATIONS for 10 variables, and 50 of the cases have missing data for at least one of the variables. In this instance, the analysis will be performed on the remaining 150 cases. Compare with PAIRWISE deletion.

—Alan Bryman

See also DELETION

LITERATURE REVIEW

A literature review is a summary of what is currently known about some issue or field on the basis of research evidence, and/or of what lines of argument there are in relation to that issue or field. Sometimes reviews are designed to stand alone, perhaps even being substantial book-length pieces of work. More usually, they amount to chapters in monographs or theses; and here their function is to set the scene for the particular study being reported, showing how it fits into the existing literature.

In recent years, some disagreement has emerged about what form literature reviews ought to take. The most common form is what has come to be labeled as *narrative* review, in which the findings and methods of relevant studies are outlined and discussed with a view to presenting an argument about the conclusions that can be drawn from the current state of knowledge in a field. Such reviews can be contrasted with annotated bibliographies and with so-called “systematic” reviews. Recently, there have also been proposals for “interpretive” reviews.

In pure form, an annotated bibliography lists studies, perhaps in alphabetical order, and provides a brief summary of each, with or without an evaluation of their strengths and weaknesses, the validity of their findings, and so forth. There is, however, a spectrum between this and the traditional literature review, in terms of how far the discussion is subordinated to an overall narrative line.

SYSTEMATIC REVIEWING has become quite influential in recent times, notably because it is seen as providing the basis for evidence-informed policymaking and practice (see Cooper, 1998; Light & Pillemer, 1984; Slavin, 1986). Systematic reviews have a number of distinctive features, frequently including the following:

- They employ explicit procedures to determine what studies are relevant.
- They rely on an explicit hierarchy of research designs according to the likely validity of the findings these produce. Studies are included or excluded on the basis of whether they employed a sufficiently strong research design, or in terms of which offer the best evidence available. Generally speaking, RANDOMIZED CONTROLLED TRIALS are at the top of the list.
- They seek to include in the review *all* studies that meet the specified criteria of relevance and validity.
- They aim to synthesize the evidence produced by these studies, rather than simply summarizing the findings of each study and relating these in a narrative or interpretive way. Systematic reviews often focus primarily on quantitative studies and employ statistical META-ANALYSIS for the purpose of synthesis.

In many ways, systematic reviewing is an attempt to apply the canons of QUANTITATIVE methodology to the process of reviewing itself (see Hammersley, 2001).

By contrast, interpretive reviewing involves application of a QUALITATIVE methodological approach. Here, the emphasis is on the interpretative role of the reviewer in making sense of the findings of different studies to construct a holistic picture of the field, a picture that may well reflect the particular interests and sensibilities of the reviewer (see Eisenhart, 1998; Schwandt, 1998). This is a development out of the traditional or narrative review.

In addition, some commentators have proposed what we might refer to as POSTMODERNIST literature reviews (see Lather, 1999; Livingston, 1999; Meacham, 1998). Here, any notion of an authoritative

account of current knowledge is rejected on the grounds that the claim to true representation of the world, at the level of particular studies or of a review, can be no more than the imposition of one discursive construction over others; thereby reflecting, for example, the work of power or of resistance to it.

Whatever approach to reviewing is adopted, certain decisions have to be made concerning the following:

- Who is the intended readership?
- How is the review to be structured?
- How are relevant studies to be found, and which studies are to be included?
- How much detail is to be provided about each study discussed; in particular, how much information is to be given about the research methods employed?
- How are the studies and their findings to be evaluated and related to one another?

In many ways, the issue of audience is the primary one because this will shape answers to the other questions, in terms of what background knowledge can be assumed, and what readers will have the greatest interest in. Recently, considerable emphasis has been put on literature reviews designed for “users” of research—in other words, on the literature review as a bridge by which research can be made relevant to policymakers, professional practitioners of one sort or another, and relevant publics. However, literature reviews also play an important function within research communities: in defining and redefining an intellectual field, giving researchers a sense of the collective enterprise in which they are participating, and enabling them to coordinate their work with one another.

—Martyn Hammersley

REFERENCES

- Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.
- Eisenhart, M. (1998). On the subject of interpretive reviews. *Review of Educational Research*, 68(4), 391–399.
- Hammersley, M. (2001). On “systematic” reviews of research literatures: A “narrative” reply to Evans and Benefield. *British Educational Research Journal*, 27(5), 543–554.
- Hart, C. (1998). *Doing a literature review*. London: Sage.
- Hart, C. (2001). *Doing a literature search*. London: Sage.
- Lather, P. (1999). To be of use: The work of reviewing. *Review of Educational Research*, 69(1), 2–7.

- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
- Livingston, G. (1999). Beyond watching over established ways: A review as recasting the literature, recasting the lived. *Review of Educational Research*, 69(1), 9–19.
- Meacham, S. J. (1998). Threads of a new language: A response to Eisenhart's "On the subject of interpretive reviews." *Review of Educational Research*, 68(4), 401–407.
- Schwandt, T. A. (1998). The interpretive review of educational matters: Is there any other kind? *Review of Educational Research*, 68(4), 409–412.
- Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, 15(9), 5–11.

LIVED EXPERIENCE

Human experience is the main epistemological basis for QUALITATIVE RESEARCH, but the concept of "lived experience" (translated from the German *Erlebnis*) possesses special methodological significance. The notion of "lived experience," as used in the works of Dilthey (1985), Husserl (1970), Merleau-Ponty (1962), and their contemporary exponents, announces the intent to explore *directly* the originary or prereflective dimensions of human existence. The etymology of the English term *experience* does not include the meaning of *lived*—it derives from the Latin *experientia*, meaning "trial, proof, experiment, experience." But the German word for experience, *Erlebnis*, already contains the word *Leben*, "life" or "to live." The verb *erleben* literally means "living through something."

Wilhelm Dilthey (1985) offers the first systematic explication of lived experience and its relevance for the human sciences. He describes "lived experience" as a reflexive or self-given awareness that inheres in the temporality of consciousness of life as we live it. "Only in thought does it become objective," says Dilthey (p. 223). He suggests that our language can be seen as an immense linguistic map that names the possibilities of human lived experiences.

Edmund Husserl (1970) uses the concept of *Erlebnis* alongside *Erfahrung*, which expresses the full-fledged acts of consciousness in which meaning is given in intentional experiences. "All knowledge 'begins with experience' but it does not therefore 'arise' from experience," says Husserl (p. 109).

Erfahrungen (as meaningful lived experiences) can have a transformative effect on our being. More generally, experience can be seen in a passive modality as something that befalls us, overwhelms us, or strikes us, and it can be understood more actively as an act of consciousness in appropriating the meaning of some aspect of the world. Thus we can speak of an "experienced" person when referring to his or her mature wisdom, as a result of life's accumulated meaningful and reflective experiences.

In *Truth and Method*, Hans-Georg Gadamer (1975) suggests that there are two dimensions of meaning to lived experience: the immediacy of experience and the content of what is experienced. Both meanings have methodological significance for qualitative inquiry. They refer to "the immediacy with which something is grasped and which precedes all interpretation, reworking, and communication" (p. 61). Thus, lived experience forms the starting point for inquiry, reflection, and interpretation. This thought is also expressed in the well-known line from Merleau-Ponty (1962), "The world is not what I think, but what I live through" (pp. xvi–xvii). If one wants to study the world as lived through, one has to start with a "direct description of our experience as it is" (p. vii).

In contemporary human science, "lived experience" remains a central methodological notion (see van Manen, 1997) that aims to provide concrete insights into the qualitative meanings of phenomena in people's lives. For example, medical science provides for strategies of intervention and action based on diagnostic and prognostic research models, but predictions about the clinical path of an illness do not tell us how (different) people actually experience their illness. PHENOMENOLOGICAL human science investigates lived experience to explore concrete dimensions of meaning of an illness, such as multiple sclerosis (Toombs, 2001) or pain in the context of clinical practice (Madjar, 2001).

Even in the DECONSTRUCTIVE and POSTMODERN work of the more language-oriented scholars such as Jacques Derrida, the idea of lived experience reverberates in phrases such as the "singularity of experience" or "absolute existence." Derrida speaks of this originary experience as "the resistance of existence to the concept or system" and says, "this is something I am always ready to stand up for" (Derrida & Ferraris, 2001, p. 40). In the human sciences, the focus on experience remains so prominent because of its

power to crack the constraints of conceptualizations, codifications, and categorical calculations. In addition, the critical questioning of the meaning and ultimate source of lived experience ensures an openness that is a condition for discovering what can be thought and found to lie beyond it.

—Max van Manen

REFERENCES

- Derrida, J., & Ferraris, F. (2001). *A taste for the secret*. Cambridge, UK: Polity.
- Dilthey, W. (1985). *Poetry and experience: Selected works* (Vol. 5). Princeton, NJ: Princeton University Press.
- Gadamer, H.-G. (1975). *Truth and method*. New York: Seabury.
- Husserl, E. (1970). *Logical investigations* (Vols. 1–2). London: Routledge & Kegan Paul.
- Madjar, I. (2001). The lived experience of pain in the context of clinical practice. In S. K. Toombs (Ed.), *Handbook of phenomenology and medicine* (pp. 263–277). Boston: Kluwer Academic.
- Merleau-Ponty, M. (1962). *Phenomenology of perception*. London: Routledge & Kegan Paul.
- Toombs, S. K. (2001). Reflections on bodily change: The lived experience of disability. In S. K. Toombs (Ed.), *Handbook of phenomenology and medicine* (pp. 247–261). Boston: Kluwer Academic.
- van Manen, M. (1997). *Researching lived experience: Human science for an action sensitive pedagogy*. London, Ontario: Althouse.

LOCAL INDEPENDENCE

Local independence is the basic assumption underlying LATENT VARIABLE models, such as FACTOR ANALYSIS, LATENT TRAIT ANALYSIS, LATENT CLASS ANALYSIS, and LATENT PROFILE ANALYSIS. The axiom of local independence states that the observed items are independent of each other, given an individual's score on the latent variable(s). This definition is a mathematical way of stating that the latent variable explains why the observed items are related to one another.

We will explain the principle of local independence using an example taken from Lazarsfeld and Henry (1968). Suppose that a group of 1,000 persons are asked whether they have read the last issue of magazines *A* and *B*. Their responses are as follows:

Table 1

	<i>Read A</i>	<i>Did Not Read A</i>	<i>Total</i>
Read B	260	240	500
Did not read B	140	360	500
Total	400	600	1,000

It can easily be verified that the two variables are quite strongly related to one another. Readers of *A* tend to read *B* more often (65%) than do nonreaders (40%). The value of phi coefficient, the product-moment correlation coefficient in a 2×2 table, equals 0.245.

Suppose that we have information on the respondents' educational levels, dichotomized as high and low. When the 1,000 people are divided into these two groups, readership is observed in Table 2. Thus, in each of the 2×2 tables in which education is held constant, there is no association between the two magazines: the phi coefficient equals 0 in both tables. That is, reading *A* and *B* is independent within educational levels.

The reading behavior is, however, very different in the two subgroups: the high-education group has much higher probabilities of reading for both magazines (0.60 and 0.80) than the low-education group (0.20 and 0.20). The association between *A* and *B* can thus fully be explained from the dependence of *A* and *B* on the third factor, education. Note that the cell entries in the marginal *AB* table, $N(AB)$, can be obtained by the following:

$$N(AB) = N(H) \cdot P(A|H) \cdot P(B|H) \\ + N(L) \cdot P(A|L) \cdot P(B|L),$$

where *H* and *L* denotes high and low education, respectively.

In latent variable models, the observed variables are assumed to be locally independent given the latent variable(s). This means that the latent variables have the same role as education in the example. It should be noted that this assumption not only makes sense from a substantive point of view, but it is also necessary to identify the unobserved factors.

The fact that local independence implies that the latent variables should fully account for the associations between the observed items suggests a simple test of this assumption. The estimated two-way tables according to the model, computed as shown above for

Table 2

	High Education			Low Education		
	Read A	Did Not Read A	Total	Read A	Did Not Read A	Total
Read B	240	60	300	20	80	100
Did not read B	160	40	200	80	320	400
Total	400	100	500	100	400	500

$N(AB)$, should be similar to the observed two-way tables. In the case of continuous indicators, we can compare estimated with observed correlations.

Variants of the various types of latent variable models have been developed that make it possible to relax the local independence assumption for certain pairs of variables. Depending on the scale type of the indicators, this is accomplished by inclusion of direct effects or correlated errors in the model.

—Jeroen K. Vermunt and Jay Magidson

REFERENCES

- Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Arnold.
- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

LOCAL REGRESSION

Local regression is a strategy for fitting smooth curves to BIVARIATE or MULTIVARIATE data. Traditional strategies (such as ordinary least squares [OLS] REGRESSION analysis) are *parametric*, in that they require the analyst to specify the FUNCTIONAL form of Y 's dependence on X prior to fitting the curve. In contrast, local regression is NONPARAMETRIC because the ALGORITHM tries to follow the empirical concentration of data points, regardless of the shape of the curve that is required to do so. The smooth curve is fitted without any advance specification of the functional relationship between the variables.

The basic principles underlying local regression have been known since the 19th century. However, it is a computationally intensive analytic strategy, so major developments did not really occur until the late 1970s and early 1980s. Currently, the most popular local

regression algorithm is the loess procedure, developed by William S. Cleveland and his associates. The term *loess* (or *lowess*, as it was previously known) is an acronym for *locally weighted regression*.

FITTING A LOESS CURVE TO DATA

Assume that the DATA consist of n observations on two variables, X and Y . These data can be displayed in a bivariate SCATTERPLOT, with X values arrayed along the horizontal axis and Y values along the vertical axis. The horizontal axis of the scatterplot also includes a set of m locations, v_j , where j runs from 1 to m . These v_j s are equally spaced across the entire range of X values.

Loess uses a “vertical sliding window” that moves across the horizontal scale axis of the scatterplot. The window stops and estimates a separate regression equation (using WEIGHTED LEAST SQUARES) at each of the m different v_j s. The width of this window is adjusted so that it always contains a prespecified proportion (often called α) of the total data points. In this sense, each of the m regressions is “local” because it only incorporates the subset of αn observations that falls within the current window. Furthermore, observations included within each local regression are weighted inversely according to their distance from the current v_j ; accordingly, observations closer to v_j have more influence on the placement of the local regression line than observations that fall farther away within the current window.

The coefficients from each local regression are used to estimate a predicted value, designated $\hat{g}(v_j)$, for the current window. The m ordered pairs, $(v_j, \hat{g}(v_j))$, are plotted in the scatterplot, superimposed over the n data points. Finally, adjacent fitted points are connected by line segments. The v_j s are located relatively close to each other along the horizontal axis, so the series of connected line segments actually appears to be a

smooth curve passing through the data points. Visual inspection of the resultant loess curve is used for substantive interpretation of the relationship between X and Y .

The general principles of multivariate loess are identical to the bivariate case. However, the fitting windows, the distances between the v_j s and the observations, and the local weights are now calculated within the k -dimensional subspace generated by the k independent variables. The end result of a multivariate loess analysis is a smooth surface that tends to coincide with the center of the data point cloud throughout the entire $(k + 1)$ dimensional space formed by the independent variables and the dependent variable.

LOESS-FITTING PARAMETERS

Loess is a nonparametric fitting procedure. However, there are still some parameters that the analyst must specify prior to the analysis. Selecting the values for these parameters can be a subjective process, but the considerations involved are quite straightforward.

First, there is α , the smoothing parameter. This gives the size of the local fitting window, defined as the proportion of observations included within each local regression. Larger α values produce smoother loess curves. The usual objective is to find the largest α value that still produces a curve that captures the major features within the data (i.e., passes through the most dense regions of the data points).

Second, the analyst must specify whether the local regressions are LINEAR or QUADRATIC in form. This decision is usually based on visual inspection of the data. If the data exhibit a generally MONOTONIC relationship between Y and X , then local linear fitting is sufficient. If there are nonmonotonic patterns, with local minima and/or maxima, then quadratic local regressions provide the flexibility for the fitted surface to conform to the “undulations” in the data.

Third, the analyst must specify whether robustness corrections are to be included in each local regression. Because a relatively small number of data points are included in each local regression, a few discrepant observations can have an adverse impact on the coefficient estimates. The robustness step down-weights the influence of any outlying observations to make sure that the fitted loess curve follows the most heavily populated regions within the data space.

CONCLUSION

The great advantage of local regression methods, such as loess, is that they are very flexible about the exact nature of the relationship between the variables under investigation. Therefore, local regression is an important tool for EXPLORATORY DATA ANALYSIS. In addition, INFERENTIAL procedures have also been developed for the methodology, making it very useful for characterizing NONLINEAR structures that may exist within a data set. The major weakness is that these methods produce graphical, rather than numerical, output—that is, the fitted smooth curve or surface itself. Nonparametric fitting strategies cannot be used to characterize the data in terms of a simple equation. Nevertheless, local regression lets the data “speak for themselves” to a greater extent than traditional parametric methods such as linear regression analysis.

—William G. Jacoby

REFERENCES

- Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596–610.
- Fox, J. (2000a). *Multiple and generalized nonparametric regression*. Thousand Oaks, CA: Sage.
- Fox, J. (2000b). *Nonparametric simple regression*. Thousand Oaks, CA: Sage.
- Jacoby, W. G. (1997). *Statistical graphics for univariate and bivariate data*. Thousand Oaks, CA: Sage.
- Loader, C. (1999). *Local regression and likelihood*. New York: Springer.

LOESS. See LOCAL REGRESSION

LOGARITHM

A logarithm is an exponent. It is the power to which a base—usually 10 or e —must be raised to produce a specific number. Thus, in base 10, the logarithm of 1,000 would be 3 because $10^3 = 1,000$. Symbolically, this relationship is also expressed as $\log_{10} 1,000 = 3$. Logarithm FUNCTIONS transform variables from original units into logarithmic units. In the example above, 1,000 is transformed into the base-10 logarithmic unit

Table 1 The Base-10 Logarithm Function and the Relationship Between x and y

x	$y = \log_{10} x$	$10^y = x$	<i>A One-Unit change in $y = a$ 90% Change in x</i>
1	0	$10^0 = 1$	
10	1	$10^1 = 10$	$\{(10 - 1)/1\} \times 100 = 90\%$
100	2	$10^2 = 100$	$\{(100 - 10)/10\} \times 100 = 90\%$
1,000	3	$10^3 = 1,000$	$\{(1,000 - 100)/100\} \times 100 = 90\%$
10,000	4	$10^4 = 10,000$	$\{(10,000 - 1,000)/1,000\} \times 100 = 90\%$
100,000	5	$10^5 = 100,000$	$\{(100,000 - 10,000)/10,000\} \times 100 = 90\%$

of 3. Of the two bases used most frequently (base 10 and base e), the natural logarithm function with the base of e , or approximately 2.7182, is often used for transforming variables to include in REGRESSION MODELS. However, both the common logarithm function (base 10) and the natural logarithm function (base e) are discussed below and demonstrate the properties of logarithms in general.

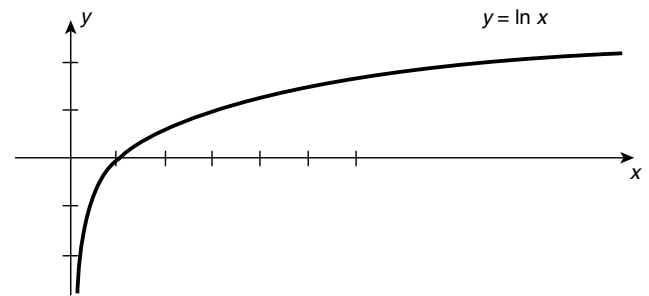
The common logarithm function transforms x from original units into y , in base-10 logarithmic units. Again, this is denoted $y = \log_{10} x$ or, equivalently, $10^y = x$. The relationship between values of x (in original units) and y (in base-10 logarithmic units) for the base-10 logarithm function is demonstrated in Table 1. Values of x increase by multiples of 10 as values of y increase by 1. So as x gets larger, it requires larger increases to produce a one-unit change in y . For example, as x goes from 10 to 100, y increases by 1, and as x goes from 1,000 to 10,000, y still increases by only 1. This demonstrates a property common to all logarithms, regardless of base; that is, the resulting scale of y shrinks gaps between values of x that are larger than 1. In practice, this “shrinking” effect pulls in large positive values and sometimes prevents having to throw away OUTLIERS that otherwise might complicate regression modeling efforts. Hence, logarithm functions are often used to transform variables with extreme positive values into variables with DISTRIBUTIONS that are more symmetric.

It might seem that including logarithmic TRANSFORMATIONS in regression models would make them difficult to interpret. However, because of the property of logarithms demonstrated in the last column of Table 1, this is not necessarily the case. For a base-10 logarithm, each unit increment in y (logarithmic units) is associated with a constant 90% increase in x (original units). For example, when a base-10 logarithm is used

as an EXPLANATORY VARIABLE in a regression model, the corresponding REGRESSION COEFFICIENT can be interpreted as the expected change in the outcome variable, given a unit change in the explanatory variable (logarithmic units) or a 90% change in the explanatory variable (original units).

The natural logarithm function (using base e) demonstrates other general properties of logarithms. Although logarithmic transformations diminish differences at large positive values of x , they accentuate differences at values of x greater than 0 and less than 1. This property is demonstrated in Figure 1, a graph of the natural logarithm function, denoted either or $y = \ln x$ or $y = \log_e x$. As x takes on smaller fractional values, asymptotically approaching 0, y becomes increasingly negative at an exponential rate. This “stretching” property is why logarithmic functions are used to transform variables, with compression in the region between 0 and 1, into variables that are distributed more like a NORMAL DISTRIBUTION. Figure 1 also demonstrates that the natural logarithm function, like all logarithm functions, is not defined for negative x or for x equal to 0.

—Jani S. Little

**Figure 1** The Natural Logarithm Function

REFERENCES

- Hamilton, L. C. (1990). *Modern data analysis: A first course in applied statistics*. Pacific Grove, CA: Brooks/Cole.
- Hamilton, L. C. (1992). *Regression with graphics: A second course in applied statistics*. Pacific Grove, CA: Brooks/Cole.
- Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.
- Swokowski, E. W. (1979). *Calculus with analytic geometry* (2nd ed.). Boston: Prindle, Weber & Schmidt.

LOGIC MODEL

A logic model depicts the theoretical or model ASSUMPTIONS and linkages between program elements and hypothesized intermediate and long-term outcomes. The logic model is often used as a tool by program evaluators during the design phase of the research to conceptualize and guide evaluation strategies by providing a comprehensive framework for data collection and analysis. The model assumptions are often referred to as *program theory* and describe the program components and rationale, the process by which the program is intended to work, and the expected outcomes of the program. The logic model encompasses the key elements of the program THEORY and provides a system-level map, which identifies

the interrelationships of the program elements, and the internal and external factors that might influence the hypothesized outcomes.

The key elements of a logic model are depicted in Figure 1 and include the following:

- The program's goals
- The available resources for supporting the program
- The target POPULATIONS or people who are affected by the program
- The program components or key activities that comprise the program
- The program outputs or products that result from the program
- The intermediate and long-term outcomes of the program
- The antecedent variables or factors outside of the program that might affect the outputs of the program
- The intervening or rival events outside of the program that might affect the hypothesized short-term and long-term outcomes of the program

As described by Yin (1998), the strength of a logic model for program evaluators is its depiction of an explicit and comprehensive chain of events or

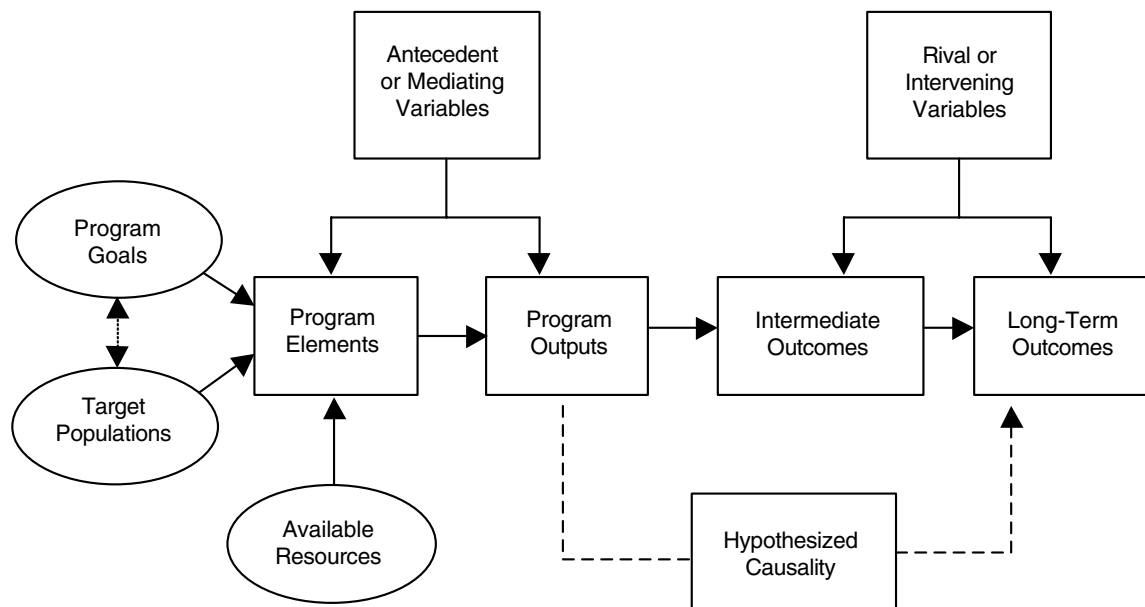


Figure 1 General Program Evaluation Logic Model

interrelationships sufficiently detailed for operational MEASURES to be determined for each step in the chain. The operational measures are used by evaluators as part of the data collection strategy (i.e., which quantitative or qualitative data need to be collected to be able to comprehensively measure the program outputs or outcomes?) and the analysis strategy (i.e., what mediating and rival events might affect program outcomes?).

In a comprehensive logic model, operational measures can include data from multiple sources that can be converged for additional clarity and insight into each stage of the evaluation. For example, an assessment of the intermediate outcomes of the program outputs can vary by target population or stakeholder perspective. Similarly, antecedent or rival variables may affect outputs and outcomes differently for different target populations. Multiple perspectives or inputs rather than a single input into the evaluation strategy better inform the extent to which the desired outcomes of the program have been achieved at each stage of the evaluation.

In addition to program evaluator utility, logic models can also be used by program managers to inform strategies and decisions for program design and implementation, and serve as a framework for reporting program outcomes. Logic models can be used to clarify for stakeholders what critical factors are required to achieve the desired outcomes, what sequence of events are potentially influencing the outcomes, what performance measures are most relevant for different target populations, how program outcomes vary across target populations, what factors beyond the program control influence intermediate and long-term outcomes across program participants or target populations, and what resources are required to achieve the desired outcomes.

—Georgia T. Karuntzos

REFERENCES

- Cooksy, L. J., Gill, P., & Kelly, P. A. (2001). The program logic model as an integrative framework for a multimethod evaluation. *Evaluation and Program Planning*, 24, 119–128.
- Julian, D. A. (1997). The utilization of the logic model as a system level planning and evaluation device. *Evaluation and Program Planning*, 20(3), 251–257.
- Millar, A., Simeone, R. S., & Carnevale, J. T. (2001). Logic models: A system tool for performance management. *Evaluation and Program Planning*, 24, 73–81.
- Yin, R. K. (1998). The abridged version of case study research. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 229–259). Thousand Oaks, CA: Sage.

LOGICAL EMPIRICISM. See LOGICAL POSITIVISM

LOGICAL POSITIVISM

Logical positivism, sometimes also called logical empiricism, is an approach to philosophy that developed in the 1920s and remained very influential until the 1950s. It was inspired by rapid and dramatic developments that had taken place in physical science and mathematics. For logical positivists, the main task was to analyze the structure of scientific reasoning and knowledge. This was important because they took science to be the model for intellectual and social progress.

Logical positivism combined EMPIRICISM with the new mathematical logic pioneered by Russell, Frege, and others. Its central doctrine was that all meaningful statements are either empirical or logical in character. This was initially interpreted to mean that they are either open to test by observational evidence or are matters of definition and therefore tautological—a principle that was referred to as the verifiability principle. It ruled out, and was designed to rule out, a great amount of philosophical and theological discussion as literally meaningless. By applying this principle, it was hoped to identify those problems that are genuine ones, as opposed to those that are simply generated by the misuse of language. Not only statements about God or the Absolute but also those about a “real” world beyond experience were often dismissed.

One of the key ideas of logical positivism was a distinction between the context of discovery and the context of justification, between how scientists develop their ideas and what is involved in demonstrating the validity of scientific conclusions. The logical positivists concentrated exclusively on the latter. Moreover, they assumed the unity of science in methodological terms: They denied that there are any fundamental differences between the natural and the social sciences. Indeed, they saw philosophy as itself part of science.

Logical positivist ideas were developed by the Vienna and Berlin Circles, which were small groupings of philosophers, mathematicians, and scientists

meeting together in the 1920s and 1930s. The Vienna Circle was the more influential of the two, producing a manifesto and running its own journal. Despite the focus on natural science, several members also had a close interest in social and political issues. Indeed one, Otto Neurath, was a sociologist and became a socialist politician for a short period. The logical positivists regarded their philosophy as capable of playing an important role in undercutting the influence of religion and other sources of irrational ideas. Two other important 20th-century philosophers, Wittgenstein and Popper, had close associations with the Vienna Circle—indeed, Wittgenstein's *Tractatus* was a significant early influence on it—although both were also critics of logical positivism, albeit for different reasons. In the 1930s, key figures in the Vienna Circle began to emigrate to Britain and the United States as a result of the rise of fascism in continental Europe, and this facilitated the spread of logical positivism across the Anglo-American world, as did A. J. Ayer's (1936) best-selling book *Language, Truth and Logic*.

Logical positivism was never a unified body of ideas, and its influence on the development of social science methodology was probably less than is often supposed; although the related doctrine of OPERATIONALISM certainly did have a strong influence on psychology and, to a lesser extent, on sociology in the first half of the 20th century. In social science, these ideas came into conflict with older views about the nature of science, as well as with the notion that the study of physical and of social reality must necessarily take different forms.

Logical positivists themselves recognized problems with their initial positions and deployed considerable effort and intelligence in trying to deal with these. One problem concerned the epistemic status of the verifiability principle itself: Was it to be interpreted as an empirical statement open to observational test, or was it a tautological matter of definition? Because it did not seem to fit easily into either category, this principle came to be judged as self-refuting by some commentators.

Another problem arose from the logical positivists' empiricism: their assumption that there can be non-inferential empirical givens, such as direct observations, which can provide an indubitable foundation for knowledge. It came to be recognized that any observation, as soon as it is formulated, relies on assumptions,

many of which may not have been tested and none of which can be tested without reliance on further assumptions. As a result, the sharp distinction that the positivists wanted to make between theoretical and empirical statements could not be sustained.

The logical positivists also failed to develop a satisfactory solution to the problem of INDUCTION: of how universal theoretical conclusions could be derived from particular empirical data with the same level of logical certainty as deductive argument. Indeed, it became clear that any body of data is open to multiple theoretical interpretations, and this threatened to make scientific theory itself meaningless in terms of the verifiability principle. Given this, the principle came to be formulated in more qualified terms. As a result, it began to provide a more realistic account of science, but the changes endangered any neat demarcation from what the logical positivists had dismissed as meaningless metaphysics.

Logical positivists dominated the philosophy of science until the 1950s, when the problems facing their position became widely viewed as irresolvable. In response, some philosophers (e.g., Quine) retained the spirit of the position while abandoning key elements of it. Others adopted competing accounts of the philosophy of science, such as that of Popper, and sought to build on these—for instance, Lakatos and Feyerabend. There were also some, such as Rorty, who reacted by drawing on pragmatism, on the work of Wittgenstein, and on continental philosophical movements. Radical responses to the collapse of logical positivism sometimes involved reduction of the philosophy of science to its history or sociology, rejection of natural science as a model, and even abandonment of the whole idea that a rational method of inquiry can be identified.

—Martyn Hammersley

REFERENCES

- Achinstein, P., & Barker, S. F. (1969). *The legacy of logical positivism*. Baltimore: Johns Hopkins University Press.
- Ayer, A. J. (1936). *Language, truth and logic*. London: Victor Gollancz.
- Friedman, M. (1991). The re-evaluation of logical positivism. *Journal of Philosophy*, 88(10), 505–519.
- Passmore, J. (1967). Logical positivism. In P. Edwards (Ed.), *The encyclopedia of philosophy* (Vol. 5, pp. 52–57). New York: Macmillan.
- Reichenbach, H. (1951). *The rise of scientific philosophy*. Berkeley: University of California Press.

LOGISTIC REGRESSION

The situation in which the outcome variable (Y) is limited to two discrete values—an event occurring or not occurring, with a “yes” or “no” response to an attitude question, or a characteristic being present or absent, usually coded as 1 or 0 (or 1 or 2), respectively—is quite typical in many areas of social science research, as in dropping out of high school, getting a job, joining a union, giving birth to a child, voting for a certain candidate, or favoring the death penalty. Logistic regression is a common tool for statistically modeling these types of discrete outcomes.

The overall framing and many of the features of the technique are largely analogous to ordinary least squares (OLS) REGRESSION, especially the linear probability model. Many of the warnings and recommendations, including the diagnostics, made for OLS regression apply as well. It is, however, based on a different, less stringent, set of statistical assumptions to explicitly take into account the limited nature of the outcome variable and the problems resulting from it—because the outcome variable as a function of independent variables can take only two values; the ERROR terms are not continuous, homoskedastic, or normally distributed; and the predicted probabilities are not constrained to behave linearly and not to be greater than 1 and less than 0.

Logistic regression fits a special s -shaped curve by taking a linear combination of the explanatory variables and transforming it by a logistic function (cf. PROBIT ANALYSIS) as follows:

$$\ln[p/(1 - p)] = a + BX + e,$$

where p is the probability that the value of the outcome variable is 1, and BX stands for $b_1x_1 + b_2x_2 + \dots + b_nx_n$. The model hence estimates the LOGIT, which is the natural log of the odds of the outcome variable being equal to 1 (i.e., the event occurring, the response being affirmative, or the characteristic being present) or 0 (i.e., the event not occurring, the response being negative, or the characteristic being absent). The probability that the outcome variable is equal to 1 is then the following:

$$p = [\exp(a + BX)]/[1 + \exp(a + BX)]$$

$$\text{or } p = 1/[1 + \exp(-a - BX)].$$

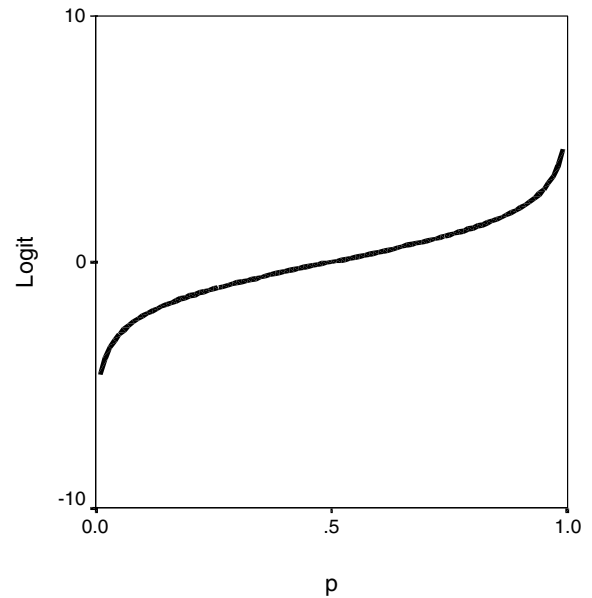


Figure 1 The Relationship Between p and Its Logit

Note that there is a nonlinear relationship between p and its logit, which is illustrated in Figure 1: In the mid-range of p there is a (near) linear relationship, but as p approaches the extremes—0 or 1—the relationship becomes nonlinear with increasingly larger changes in logit for the same change in p .

For large samples, the parameters estimated by MAXIMUM LIKELIHOOD ESTIMATION (MLE) are unbiased, efficient, and normally distributed, thus allowing statistical tests of significance. The technique can be extended to situations involving outcome variables with three or more categories (polytomous, or multinomial dependent variables).

EXAMPLE

To illustrate its application and the interpretation of the results, we use the data on the attitude toward capital punishment—“Do you favor or oppose the death penalty for persons convicted of murder?”—from the General Social Survey. Out of 2,599 valid responses in 1998, 1,906 favored (73.3%) and 693 opposed (26.7%) it. The analysis below examines the effects of race (White = 1 and non-White = 0) and education on the DEPENDENT VARIABLE, which is coded as 1 for those who favor it and 0 for those who oppose it, controlling for sex (male = 1 and female = 0) and age.

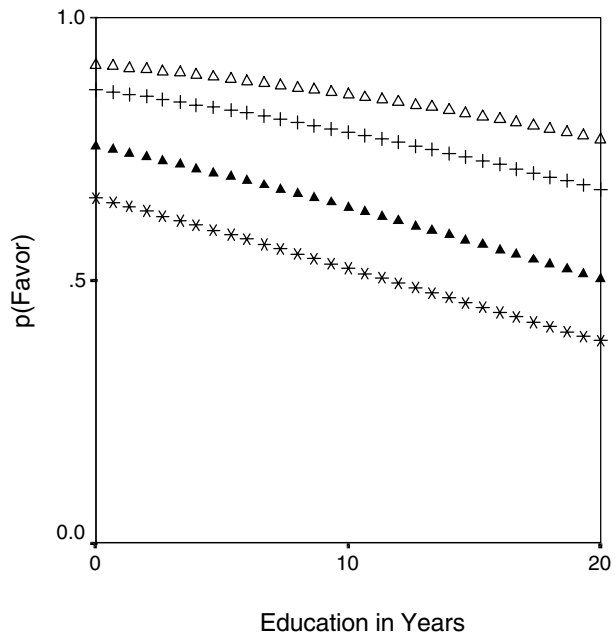
Table 1 Classification Table

Observed	Predicted		% Correct
	Oppose (0)	Favor (1)	
Oppose (0)	83	605	12.1
Favor (1)	86	1,811	95.5
Overall %			73.3

The standard logistic regression output consists of two parts: one that shows the overall fit of the model and the other that reports the parameter estimates. In the first part, several statistics can be used for comparing alternative models or evaluating the performance of a single model.

1. The model LIKELIHOOD RATIO (LR) STATISTIC, G^2 or L^2 , is used to determine if the overall model is statistically significant vis-à-vis the restricted model, which includes only a constant term. It is distributed χ^2 with k degrees of freedom, where k is the number of INDEPENDENT VARIABLES. In the example at hand, we obtain $G^2 = 156.59$, with $df = 4$ and $p < .000$. This model chi-square shows that the prediction model significantly improves on the restricted model.
2. Another method of gauging the performance of the logistic regression model involves the cross-classification of observed and predicted outcomes and summarizes the correspondence between the two with the percentage of correct predictions. For our example, Table 1 shows that the model predicts 73.3% of the cases correctly. In general, the larger the value, the better the model fit. Yet, these results depend a great deal on the initial distribution of the outcome variable.
3. Other statistics—pseudo- R^2 s—assessing the overall fit of the model are similar to R^2 in OLS, although they cannot be used for hypothesis testing.

The estimated coefficients are usually provided in the second section of the output in a format similar to that in OLS (see Table 2). The Wald statistic, which is

**Figure 2** Favoring Death Penalty by Education, Race, and Sex

distributed as χ^2 with 1 degree of freedom, is used to test the hypothesis that the coefficient is significantly different from zero. In our death penalty example, all the independent variables included in the model are significant at the .05 level. The interpretation of the coefficients is fundamentally different, though. The logistic regression coefficients indicate the change in the logit for each unit change in the independent variable, controlling for the other variables in the model. For example, the coefficient for education, $-.056$, indicates that an additional year of schooling completed reduces the logit by $.056$. Another, more intuitive, approach is to take the antilog of (i.e., to exponentiate) the coefficient, which gives the statistical effect of the independent variable in terms of odds ratios. It is multiplicative in that it indicates no effect when it has a value of 1, a positive effect when it is greater than 1, and a negative effect when it is less than 1. Each additional year of schooling, for instance, decreases the odds of favoring the death penalty by a factor of $.946$ when the other variables are controlled for. Still another approach is to calculate predicted probabilities for a set of values of the independent variables. Figure 2 shows the effect of education on the probability of favoring

Table 2 Logistic Regression Coefficient Estimates

<i>Independent Variable</i>	<i>B</i>	<i>SE</i>	<i>Wald Statistic</i>	<i>df</i>	<i>p</i>	<i>Exp</i>
Education	-.056	.016	11.586	1	.001	.946
Age	-.006	.003	4.348	1	.037	.994
Sex (male = 1)	.481	.095	25.843	1	.000	1.617
Race (White = 1)	1.190	.107	124.170	1	.000	3.289
Constant	.922	.271	11.577	1	.001	2.514

the death penalty. Separate lines are generated for non-White female (*), non-White male (▲), White female (+), and White male (Δ), with age fixed at the sample mean.

—Shin-Kap Han and C. Gray Swicegood

REFERENCES

- Liao, T. F. (1994). *Interpreting probability models: Logit, probit, and other generalized linear models*. Thousand Oaks, CA: Sage.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.
- Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.

LOGIT

The *logit* is the natural logarithm of the *odds* that an event occurs or a state exists. For example, if $P(Y)$ is the probability that someone is a computer user or that someone commits a crime, then $\text{logit}(Y) = \ln[P(Y)/(1 - P(Y))]$. The logit is negative whenever $P(Y)$ is less than .5, and is negative and infinitely large in magnitude when $P(Y) = 0$; it is positive whenever $P(Y)$ is greater than .5, and positive and infinitely large in magnitude when $P(Y) = 1$. If $P(Y) = .5$, because the natural logarithm of 1 is zero, $\text{logit}(Y) = \ln[(.5)/(.5)] = \ln(1) = 0$. The logit of a dichotomous or dichotomized dependent variable is used as the dependent variable in LOGISTIC REGRESSION and LOGIT MODELS.

—Scott Menard

LOGIT MODEL

The logit model was developed as a variant of the LOG-LINEAR MODEL to analyze categorical variables, parallel to the analysis of CONTINUOUS VARIABLES with ORDINARY LEAST SQUARES regression analysis. The basic formulation of the log-linear and logit models begins with the assumption that we are dealing with a CONTINGENCY TABLE of two or more dimensions (represented by two or more variables), each of which has two or more categories, which may be ordered but are usually assumed to be unordered. The analysis of a two-way contingency table for CATEGORICAL variables is a relatively simple problem covered in basic statistics texts. When tables contain three or more categorical variables, however, it quickly becomes cumbersome to consider all the different ways of arranging the multidimensional tables in the two-dimensional space of a page.

Log-linear models are used to summarize three-way and higher-dimensional contingency tables in terms of the effects of each variable in the table on the frequency of observations in each cell of the table; they are also used to show the relationships among the variables in the table. As described by Knoke and Burke (1980, p. 11), the *general log-linear model* makes no distinction between independent and dependent variables but in effect treats all variables as dependent variables and models their mutual associations. The expected cell frequencies based on the model (e.g., F_{ijk} for a three-variable table) are modeled as a function of all the variables in the model. The *logit model* is mathematically equivalent to the log-linear model, but one variable is considered a dependent variable, and instead of modeling the expected cell frequencies, the *odds* of being in one category (instead of any of the other categories) on the dependent variable

is modeled as a function of the other (independent) variables instead of the expected frequency being modeled.

THE GENERAL LOG-LINEAR MODEL

Consider as an example a three-way contingency table in which variable A is sex (0 = male, 1 = female), variable B is race (1 = White, 2 = Black, 3 = other), and variable C is computer use, which indicates whether (0 = no, 1 = yes) the respondent uses a personal computer. The categories of each variable are indexed by $i = 0, 1$ for A , $j = 1, 2, 3$ for B , and $k = 0, 1$ for C . For a trivariate contingency table, the *saturated model* is the model that includes the effect of each variable, plus the effects of all possible INTERACTIONS among the variables in the model. Using the notation for log-linear models,

$$F_{ijk} = \eta \tau_i^A \tau_j^B \tau_k^C \tau_{ij}^{AB} \tau_{jk}^{BC} \tau_{ijk}^{ABC}, \quad (1)$$

where each term on the right-hand side of the equation is multiplied by the other terms, and

F_{ijk} = frequency of cases in cell (i, j, k) , which is expected if the model is correct.

η = geometric mean of the number of cases in each cell in the table, $\eta = (f_{010} f_{020} f_{030} f_{110} f_{120} f_{130} f_{011} f_{021} f_{031} f_{111} f_{121} f_{131})^{1/12}$.

τ_i^A = effect of A on cell frequencies (one effect for each category of A ; one category is redundant); similarly, τ_j^B and τ_k^C are the respective effects of B and C .

τ_{ij}^{AB} = effect of the interaction between A and B ; effects that occur to the extent that A and B are not independent; similarly, τ_{ik}^{AC} and τ_{jk}^{BC} are the respective interaction effects of A with C and of B with C .

τ_{ijk}^{ABC} = effect of the three-way interaction among A , B , and C .

There is no single standard for log-linear and logit model notation. Goodman (1973) presents two notations, the first of which [the notation used in

equation (1)] expresses the row and column variables as exponents to the coefficients and has come to be more frequently used in log-linear analysis. In the second notation, which is used for logit but not general log-linear models and is more similar to the OLS regression format, the row and column variables of the table are expressed as dummy variables $X_1, X_2, \dots, X_k$. A third notation, often used for log-linear but not logit models, is the fitted marginals notation (Knoke & Burke 1980), involving the use of "curly" brackets to specify effects, for example, representing the saturated model in equation (1) as $\{A\}\{B\}\{C\}\{AB\}\{AC\}\{BC\}\{ABC\}$.

In the multiplicative model of equation (1), any $\tau > 1$ results in more than the geometric mean number of cases in a cell, but $\tau < 1$ results in less than the geometric mean number of cases in the cell. When any $\tau = 1$, the effect in question has no impact on the cell frequencies (it just multiplies them by 1; this is analogous to a coefficient of zero in OLS regression). Because the saturated model includes all possible effects, including all possible interactions, it perfectly reproduces the cell frequencies, but it has no DEGREES OF FREEDOM. As effect parameters (τ) are constrained to equal 1 (in other words, they are left out of the model), we gain degrees of freedom and are able to test how well the model fits the observed data.

Taking the natural LOGARITHM of the saturated model in equation (1) results in the following equation:

$$\begin{aligned} \ln(F_{ijk}) = & \ln(\eta) + \ln(\tau_i^A) + \ln(\tau_j^B) + \ln(\tau_k^C) \\ & + \ln(\tau_{ij}^{AB}) + \ln(\tau_{ik}^{AC}) + \ln(\tau_{jk}^{BC}) + \ln(\tau_{ijk}^{ABC}) \end{aligned} \quad (2)$$

or

$$\begin{aligned} V_{ijk} = \theta = & \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} \\ & + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}, \end{aligned} \quad (3)$$

where $V_{ijk} = \ln(F_{ijk})$, $\theta = \ln(\eta)$, and each $\lambda = \ln(\tau)$ for the corresponding effect. Equations (2) and (3) convert the multiplicative model of equation (1) to an additive model that is more similar to the OLS regression model. In equation (3), the absence of an effect is indicated by $\lambda = 0$ for that effect. For example, if the interaction between A and B has no impact on the cell frequencies, $\lambda_{ij}^{AB} = 0$, corresponding to $\ln(\tau_{ij}^{AB}) = 0$ because $\tau_{ij}^{AB} = 1$ and $\ln(1) = 0$. The additive form of the model has several advantages, as described in Knoke and Burke (1980), including ease of calculating standard errors and thus of determining the statistical significance of the parameters.

THE LOGIT MODEL

In the log-linear model described in equations (1), (2), and (3), all three variables have the same conceptual status. Computer use is treated no differently from sex or race. Substantively, however, we may be interested specifically in the impact of race and sex on computer use. The logit model is a variant of the log-linear model in which we designate one variable as the dependent variable and the other variables as predictors or independent variables. For the dichotomous dependent variable C (computer use) in the saturated model in equation (1), we can express the expected ODDS that a respondent is a computer user, Ω_C , as the ratio between the expected frequency of being a computer user, F_{ij1} , and the expected frequency of not being a computer user, F_{ij0} :

$$\begin{aligned}\Omega_C &= (F_{ij1}/F_{ij0}) \\ &= (\eta \tau_i^A \tau_j^B \tau_1^C \tau_{ij}^{AB} \tau_{i1}^{AC} \tau_{j1}^{BC} \tau_{ij1}^{ABC}) \\ &\quad / (\eta \tau_i^A \tau_j^B \tau_0^C \tau_{ij}^{AB} \tau_{i0}^{AC} \tau_{j0}^{BC} \tau_{ij0}^{ABC}) \\ &= (\tau_1^C \tau_{i1}^{AC} \tau_{j1}^{BC} \tau_{ij1}^{ABC}) / (\tau_0^C \tau_{i0}^{AC} \tau_{j0}^{BC} \tau_{ij0}^{ABC}), \quad (4)\end{aligned}$$

where η , τ_i^A , τ_j^B , and τ_{ij}^{AB} cancel out of the numerator and the denominator of the equation, and all of the remaining terms include the variable C . If we take the natural logarithm of equation (4), the dependent variable becomes the log-odds of computer use, or the *logit* of the probability that the respondent is a computer user: $\ln(\Omega_C) = 2\lambda^C + 2\lambda_i^{AC} + 2\lambda_j^{BC} + 2\lambda_{ij}^{ABC}$ or, reexpressing $\beta = 2\lambda$ and $\text{logit}(C) = \ln(\Omega_C)$,

$$\begin{aligned}\text{logit}(C) &= \ln(\Omega_C) \\ &= \beta^C + \beta^{AC} + \beta^{BC} + \beta^{ABC}. \quad (5)\end{aligned}$$

In the DUMMY VARIABLE notation mentioned earlier and described or used by, among others, Goodman (1973), Hutcheson and Sofroniou (1999), and Powers and Xie (2000),

$$\begin{aligned}\text{logit}(C) &= \beta_0 + \beta_1 X_{A=1} + \beta_2 X_{B=2} + \beta_3 X_{B=3} \\ &\quad + \beta_4 X_{A=1} X_{B=2} + \beta_5 X_{A=1} X_{B=3}. \quad (6)\end{aligned}$$

The subscripts on the X variables here indicate which categories are omitted as reference categories; male ($A = 0$) is the reference category for variable A , and White ($B = 1$) is the reference category for variable B . $X_{A=1}$ is the dummy variable for being female; $X_{B=2}$ and $X_{B=3}$ are the dummy variables for

being Black and other, respectively; and the last two terms in the equation represent the interaction between sex and race (A and B). Equations (5) and (6) represent exactly the same model; they are just two different ways of showing which effects are included in the model.

It is also possible to include continuous, interval, or ratio scale predictors in the logit model (DeMaris, 1992). Sometimes, a distinction is made between logistic regression as a method for dealing with continuous predictors, as opposed to logit analysis as a method for dealing with categorical predictors, but this distinction is somewhat artificial. In form, the two models are identical, with both being able to analyze continuous as well as categorical predictors. They differ primarily in the numerical techniques used to estimate the parameters and in how they are usually presented, with the logistic regression model appearing more like the ordinary least squares regression model and the logit model appearing more like the general log-linear model.

LOGIT MODELS FOR POLYTOMOUS VARIABLES

Suppose that instead of computer use as the dependent variable, we have the nominal variable *ice cream preference*, where 1 = vanilla, 2 = chocolate, 3 = strawberry, and 4 = other. Among the several possible approaches to analyzing this nominal variable, in which there is no natural ordering to the categories, is the MULTINOMIAL LOGIT model, in which one category (e.g., 4 = other) is selected as a reference category, and each of the other categories is separately compared to the reference category. Because the reference category cannot be compared to itself, this results in a series of $k - 1 = 3$ equations, one for each category compared to the reference category, and we add subscripts to the coefficients in equation (6) to indicate which of the other categories we are contrasting with the first category:

$$\begin{aligned}\text{logit}(C_1) &= \ln[P(C = 1)/P(C = 4)] \\ &= \ln(F_{ij1}/F_{ij4}) = \beta_1^C + \beta_1^{AC} + \beta_1^{BC} + \beta_1^{ABC}, \\ \text{logit}(C_2) &= \ln[P(C = 2)/P(C = 4)] \\ &= \ln(F_{ij2}/F_{ij4}) = \beta_2^C + \beta_2^{AC} + \beta_2^{BC} + \beta_2^{ABC}, \\ \text{logit}(C_3) &= \ln[P(C = 3)/P(C = 4)] \\ &= \ln(F_{ij3}/F_{ij4}) = \beta_3^C + \beta_3^{AC} + \beta_3^{BC} + \beta_3^{ABC},\end{aligned}$$

where $P(C = 1)$ is the probability that C has the value 1, $P(C = 2)$ is the probability that C has the value 2, and so forth, given the values of the predictors, and the equations indicate how well sex and race predict preferences for vanilla [$\text{logit}(C_1)$], chocolate [$\text{logit}(C_2)$], and strawberry [$\text{logit}(C_3)$] as opposed to other flavors of ice cream. The multinomial logit model is more complicated to interpret, with $(k - 1)$ sets of coefficients, but it avoids unnecessary and artificial recoding of the dependent variable into a single dichotomy.

As another alternative, suppose that the dependent variable C represents different levels of academic attainment, with 1 = *less than high school graduation*, 2 = *high school graduation but no college degree*, 3 = *undergraduate college degree but no graduate degree*, and 4 = *graduate (master's or doctoral) degree*. We could construct a multinomial model similar to the one above, but the additional information provided by the natural ordering of the categories, from lower to higher levels of education, may allow us to construct a simpler model. If we again take the fourth category (graduate degree) as the reference category, it is plausible that the impacts of sex and race on the difference between an undergraduate degree and a graduate degree are the same as their impacts on the difference between an undergraduate degree and a high school diploma (i.e., the impacts of sex and race on the transition from one adjacent level to the next are the same regardless of which adjacent levels of education are being compared). If so, we may use the *adjacent categories logit* model, in which $\ln[P(C = k)/P(C = k+1)] = \sum \beta_k^C + \beta^{AC} + \beta^{BC} + \beta^{ABC}$ or, for the four-category dependent variable used here, $\text{logit}(C) = \beta_1^C + \beta_2^C + \beta_3^C + \beta^{AC} + \beta^{BC} + \beta^{ABC}$, and only the coefficient for β^C is subscripted to show that we have a different starting point (a *threshold*, much like the intercept in a multinomial logit model) for the logit regression line in each category. Another possibility is the *cumulative logit* model, in which the equation is basically the same, but a different logit is used as the dependent variable: $\ln[P(C \leq k)/P(C > k)] = \sum \beta_k^C + \beta^{AC} + \beta^{BC} + \beta^{ABC}$ or, again, $\text{logit}(C) = \beta_1^C + \beta_2^C + \beta_3^C + \beta^{AC} + \beta^{BC} + \beta^{ABC}$ for the four-category dependent variable $C = \text{educational attainment}$. In effect, the adjacent categories and cumulative logit models assume parallel logit regression lines, starting at a different value for each category. Another way to describe this is to say that coefficients for the relationships among the

variables are constant, but they have a different starting point for each different level of the dependent variable. Other possible logit models can be constructed to analyze ordered dependent variables, as reviewed, for instance, in Long (1997).

The (dichotomous) logit model or its variant, the LOGISTIC REGRESSION model, is one of two models widely used in the analysis of dichotomous dependent variables and appears to be more widely used than the main alternative, the PROBIT model. The multinomial logit model is probably the model most commonly used by social scientists today to analyze a polytomous nominal dependent variable with two or more predictors. The family of ordered logit models, however, represents only one of several more or less widely used approaches to the analysis of polytomous-ordered categorical dependent variables. Other options in this case, as summarized by Menard (2002), include ordered probit models, treating ordinal dependent variables as though they represented interval or ratio scaled variables, ordinal regression, and the use of structural equation models with polychoric correlations. It is also possible to test whether the slopes are parallel in an ordered logit model; if they are not, a multinomial logit model or some other model may be more appropriate than the ordered logit model, even for ordered categorical data. Broadly speaking, the family of logit models, particularly if we include the special case of logistic regression, is presently the most widely used approach to the analysis of categorical dependent variables in the social sciences.

—Scott Menard

REFERENCES

- DeMaris, A. (1992). *Logit modeling*. Newbury Park, CA: Sage.
- Goodman, L. A. (1973). Causal analysis of data from panel studies and other kinds of surveys. *American Journal of Sociology*, 78, 1135–1191.
- Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. London: Sage.
- Knoke, D., & Burke, P. J. (1980). *Log-linear models*. Beverly Hills, CA: Sage.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Powers, D. A., & Xie, Y. (2000). *Statistical methods for categorical data analysis*. San Diego: Academic Press.

LOG-LINEAR MODEL

Cross-classifications of categorical variables (CONTINGENCY TABLE) are ubiquitous in the social sciences, and log-linear models provide a powerful and flexible framework for their analysis. The log-linear model has direct analogies to the linear model routinely used to perform an ANALYSIS OF VARIANCE. There are, of course, important issues that are particular to contingency tables analysis, and those are the primary focus of this entry.

This entry begins with an exposition of the log-linear model in the context of the 2×2 contingency table, and from the basic concepts and models of this simple setting, extensions or generalizations are made to illustrate how log-linear models can be applied to applications with more than two categories per variable, more than two variables, or a one-to-one matching between the categories of two variables. The concept and representation of the ODDS RATIO are fundamental to understanding the log-linear model and its applications, and this will be demonstrated throughout.

ASSOCIATION: THE 2×2 CONTINGENCY TABLE, LOG-ODDS RATIO, AND LOG-LINEAR MODEL PARAMETERS

The correlation coefficient is the standard measure for the assessment of (linear) ASSOCIATION between two continuous variables, and the parameters in the classical LINEAR REGRESSION model are readily related to it. The odds ratio is the analogous measure in the log-linear model, as well as numerous generalizations of these models that have been developed for the modeling and analysis of associations between categorical variables. That is, log-linear model parameters have immediate interpretations in terms of log-odds, log-odds ratios, and contrasts between log-odds ratios. The analogy to linear models for analysis of variance is that the parameters in those models have interpretations in terms of means, differences in means, or mean contrasts. In addition, there are immediate connections between log-odds ratios for association and the parameters in regression models for log-odds, just as there are immediate connections between PARTIAL CORRELATIONS and REGRESSION COEFFICIENTS in ordinary linear regression.

The fundamental concepts and parameterization of the log-linear model are readily demonstrated with the consideration of the 2×2 contingency table, as well as sets of 2×2 contingency tables. Let π_{ij} denote the probability associated with the cell in row i and column j of a 2×2 contingency table, and let n_{ij} denote the corresponding observed count. For example, consider the following cross-classification of applicants' gender and admission decisions for all 12,763 applications to the 101 graduate programs at the University of California at Berkeley for 1973 (see, e.g., Agresti, 1984, pp. 71–73):

Table 1 Berkeley Graduate Admissions Data

<i>Gender</i>	<i>Admissions Decision</i>	
	<i>Yes</i>	<i>No</i>
Male	3,738	4,704
Female	1,494	2,827

In this case, $n_{11} = 3,738$, $n_{12} = 4,704$, $n_{21} = 1,494$, and $n_{22} = 2,827$. These data were presented in an analysis to examine possible sex bias in graduate admissions, as it is immediately obvious from simple calculations that more than 44% (i.e., $3,738/8,442 > 0.44$) of males were offered admission, whereas less than 35% (i.e., $1,494/4,321 < 0.35$) of females were. Log-linear models can be applied to test the hypothesis that admissions decisions were (statistically) independent of applicants' gender versus the alternative that there was some association/dependence.

A log-linear model representation of the cell probabilities π_{ij} in the 2×2 contingency table is

$$\ln(\pi_{ij}) = \lambda + \lambda_i^R + \lambda_j^C + \lambda_{ij}^{RC},$$

where “ $\ln(\)$ ” is used to represent the natural logarithm function and the λ_i^R , λ_j^C , and λ_{ij}^{RC} are row, column, and association parameters, respectively, to be estimated from the data. The model of statistical independence is the reduced model where $\lambda_{ij}^{RC} = 0$ for all pairs (i, j) . If the R (row) variable (i.e., applicants' gender) and the C (column) variable (i.e., admissions decision) are not independent of one another, then the association between R and C is represented in terms of the λ_{ij}^{RC} , and a specific contrast of these terms is readily

interpreted as the log-odds ratio for assessing the degree of association between R and C ; that is,

$$\ln\left(\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\right) = \lambda_{11}^{RC} - \lambda_{12}^{RC} - \lambda_{21}^{RC} + \lambda_{22}^{RC}.$$

Note that, just as with linear models for analysis of variance, restrictions are necessary to identify individual parameters in the log-linear model (e.g., $\lambda_1^R = \lambda_1^C = \lambda_{1j}^{RC} = \lambda_{i1}^{RC} = 0$, for all i and for all j), but the contrast $\lambda_{11}^{RC} - \lambda_{12}^{RC} - \lambda_{21}^{RC} + \lambda_{22}^{RC}$ is identified as the log-odds ratio independent of the identifying restrictions chosen for the individual λ terms.

The method of MAXIMUM LIKELIHOOD is routinely employed to estimate log-linear models and their parameters; in the case of the Berkeley admissions data, the maximum likelihood estimate of the $\ln(\pi_{ij})$ in the model, not assuming that R and C were independent, is $\ln(n_{ij}/N)$, where N is the total sample size. Hence, the maximum likelihood estimate of the log-odds ratio equals $\ln\left(\frac{3738 \times 2827}{1494 \times 4704}\right) = 0.41$, and a statistic (see the ODDS RATIO entry for details on the calculation) for testing the NULL HYPOTHESIS that the log-odds ratio equals zero (i.e., statistical independence) equals 10.52—thereby providing strong evidence ($p < .001$) for a departure from independence and hence an association.

A general approach to testing hypotheses within the log-linear model framework is to fit the model under the null hypothesis (independence, in this case) and the ALTERNATIVE HYPOTHESIS (in this case, the unrestricted model that includes the λ_{ij}^{RC} terms). The difference in the LIKELIHOOD RATIO STATISTICS (which we denote here by L^2) for the two models is a test of the null hypothesis against the alternative, and (under certain assumptions) the appropriate reference distribution is a chi-square with DEGREES OF FREEDOM (which we denote here by df) equal to the difference in the (residual) degrees of freedom for the two models. For our example, the null model has $L^2 = 112.40$, and $df = 1$. The unrestricted model in the example, because it is specified with no simplifying assumptions, has residual deviance and residual degrees of freedom both equal to 0. It follows that L^2 for testing the null hypothesis of independence against an unrestricted alternative equals 112.40, and the corresponding p -value is considerably less than .001.

Note that the square root of L^2 (i.e., $\sqrt{112.40} = 10.60$) is numerically close to the value of the test statistic based on the maximum likelihood estimate of the log-odds ratio. Indeed, the similarity between the values of the two test statistics—the first based on

a parameter estimate and its estimated standard error and the second based on the log-likelihood—is not an accident. These two methods to testing hypotheses for log-linear models (and other models estimated by maximum likelihood) are asymptotically equivalent, which intuitively means that they should be numerically very close when the observed sample size is “large.”

Log-linear models are in the family of so-called GENERALIZED LINEAR MODELS, and hence software for such models (e.g., the `glm` function in R or Splus, or `proc glm` in SAS) can be used to fit and analyze log-linear models.

The standard theory of likelihood-based inference applies to log-linear modeling estimation and testing when the sampling model for the observed cross-classification is either MULTINOMIAL, product-multinomial, or independent POISSON. Inferences are not sensitive to which of these three sampling models was employed, provided that the log-linear model estimated does not violate the sampling assumptions. For example, in the case of product-multinomial sampling, the estimated model must yield estimated cell counts that preserve the relevant multinomial totals. Further discussion of these issues can be found in the general categorical data analysis references for this entry.

MODELS FOR MORE THAN TWO VARIABLES

The development of log-linear models for the setting where there are more than two categorical variables includes both the specification of models in terms of equations and the development of associated concepts of independence. The presentation here focuses on further analysis of 1973 graduate admission data from the University of California, Berkeley. The following data incorporate a third variable (in addition to applicants' gender and admission status): major applied to. Data are presented only for what were the six largest graduate programs at the time, accounting for 4,526 applications and approximately 35% of the total.

The fullest specification of a log-linear model with three variables is

$$\begin{aligned} \ln(\pi_{ijk}) = & \lambda + \lambda_i^R + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{RC} + \lambda_{ik}^{RM} \\ & + \lambda_{jk}^{RM} + \lambda_{ijk}^{RCM}, \end{aligned}$$

Table 2 Berkeley Admissions Data for Six Majors

Major	Admissions Decision			
	Male		Female	
	Yes	No	Yes	No
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317

where the additional subscript k indexes the third variable. Adopting the notation from the previous example (i.e., R represents the applicants' gender, and C is used to denote the admissions decision), M is used to denote the major applied to. The log-odds ratio to explore the association between R and C for a specific major is

$$\ln\left(\frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}\right) = \ln(\pi_{11k}) - \ln(\pi_{12k}) \\ - \ln(\pi_{21k}) + \ln(\pi_{22k}),$$

which in turn can be written as

$$\ln\left(\frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}\right) = (\lambda_{11}^{RC} - \lambda_{12}^{RC} - \lambda_{21}^{RC} + \lambda_{22}^{RC}) \\ + (\lambda_{11k}^{RCM} - \lambda_{12k}^{RCM} - \lambda_{21k}^{RCM} + \lambda_{22k}^{RCM}).$$

It is important to note that in the latter equation, the first contrast of the λ terms does not depend on the level of M , whereas the second contrast does. Let $\theta_k^{RC(M)}$ denote the log-odds ratio for the *conditional association* between gender and admissions decision in the k th major (i.e., $\theta_k^{RC(M)} = \ln(\pi_{11k}\pi_{22k}/\pi_{12k}\pi_{21k})$), and let

$$\delta^{RC} = \lambda_{11}^{RC} - \lambda_{12}^{RC} - \lambda_{21}^{RC} + \lambda_{22}^{RC}$$

and

$$\delta_k^{RCM} = \lambda_{11k}^{RCM} - \lambda_{12k}^{RCM} - \lambda_{21k}^{RCM} + \lambda_{22k}^{RCM}.$$

It then follows that $\theta_k^{RC(M)} = \delta^{RC} + \delta_k^{RCM}$. The obvious implication is that the full model allows for the possibility that any association between R and C could be variable across the levels of M ; that is, the δ_k^{RCM} are nonzero, and hence the $\theta_k^{RC(M)}$ are variable

across the majors. On the other hand, a model that excludes the three-factor terms λ_{ijk}^{RCM} (and hence δ_k^{RCM} equals 0) is one in which the association between R and C is assumed to be the same across all majors—this is a model of *homogeneous association*, or *no-three-factor interaction*. Likewise, a model excluding both the λ_{ij}^{RC} terms and the λ_{ijk}^{RCM} terms would yield a log-odds ratio of 0 for the R – C association, given $M = k$, and hence it would be a model assuming that R and C were independent in each of the six majors. This latter model is one of *conditional independence*— R and C are independent at each level of $M = k$.

The likelihood ratio test statistic for the homogeneous association model is $L^2 = 20.20$, on five residual df (i.e., six log-odds ratios are constrained to a single value, and hence $df = 6 - 1$). The corresponding p value is approximately equal to .001, indicating some departure from homogeneous association. Fitting the heterogeneous association model (i.e., the full model) to the data yields a perfect fit to the data, and inspection of the estimates $\hat{\delta}_k^{RCM}$ (not presented) shows that association between R and C in major A was markedly different from the other five majors. Indeed, there appears to be little evidence for an association between R and C across the other majors. Therefore, consider fitting the model where δ_k^{RCM} is estimated only when $k = 1$; that is, we constrain $\delta_k^{RCM} = 0$ for all k not equal to 1. The likelihood ratio test statistic is $L^2 = 2.68$ on five residual df (again, we have estimated but one out of six possible log-odds ratios), which indicates an excellent fit to the data.

There is not space here for a more thorough analysis of the Berkeley admissions data, but it is worth noting that the two examples taken in tandem provide for a nice illustration of Simpson's paradox. That is, the association between an applicant's gender and the admissions decision that is evident when we do not consider major suggests bias against female applicants, but when we control for major, we see no overall bias toward one gender. The data do support that females were admitted at a higher rate in the department receiving the largest number of applications, which is in the opposite direction of the bias observed in the two-way table combining all majors. The paradox arises because males and females did not apply to the various majors at the same rates. That is, females were relatively more likely to apply to majors with lower admissions rates, and hence their admissions rate, when combined across majors, is lower.

Table 3 Educational Attainment Data

Respondent's Level (Years)	Oldest Brother's Level (Years)		
	< 12	12	≥ 13
< 12	16	20	16
12	15	65	76
≥ 13	26	69	629

LOG-LINEAR MODELS WITH MORE THAN TWO CATEGORIES: A TWO-WAY EXAMPLE AND REFINING OR MODIFYING THE λ_{ij}^{RC} PARAMETERS

In the general setting of an $I \times J$ cross-classification log-linear model, parameters are still interpreted in terms of log-odds ratios for association; more generally, in a multiway table, they are interpreted in terms of conditional or partial log-odds ratios. For example, in an arbitrary two-way table, contrasts of the association parameters λ_{ij}^{RC} correspond to a log-odds ratio comparing the relative odds of a pair of categories with respect to one variable across two levels of the other; that is,

$$\ln\left(\frac{\pi_{ij}\pi_{i'j'}}{\pi_{ij'}\pi_{i'j}}\right) = \lambda_{ij}^{RC} - \lambda_{ij'}^{RC} - \lambda_{i'j}^{RC} + \lambda_{i'j'}^{RC}.$$

Consider the following data from a study of educational attainment. The data are from the 1973 Occupational Changes in a Generation II Survey (Mare, 1994), and here we only consider data for the sons of fathers with 16 or more years of education. The rows correspond to the educational attainment of the survey respondents, and the columns correspond to the educational attainment of the respondents' oldest brothers. The likelihood ratio statistic for testing statistical independence between respondents' educational attainment and oldest brothers' educational attainment is $L^2 = 161.86$. The relevant number of degrees of freedom is four (i.e., $(3 - 1) \times (3 - 1) = 4$), the corresponding p value for a chi-square distribution with 4 df is much less than .001, and hence there is strong evidence indicating a departure from independence. We could fit the saturated model to estimate the four identifiable λ_{ij}^{RC} , but instead we use these data to illustrate how more parsimonious models, arrived at by placing restrictions on the λ_{ij}^{RC} , can be developed to explore specific departures from independence.

The first model we consider exploits the fact that the categories for both variables are ordered. Let

$\lambda_{ij}^{RC} = \theta ij$, and then a log-odds ratio comparing any pair of adjacent categories equals θ ; that is,

$$\ln\left(\frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}}\right) = \theta.$$

Such a model is referred to as an example of ASSOCIATION MODELS, and in the case of the example, $L^2 = 33.12$ on 3 df for this model. This is a dramatic improvement in fit, but there remains evidence of lack of fit.

Next, consider refining the model further by adding a parameter that enters the model only for observations along the main diagonal—that is, let $\lambda_{ij}^{RC} = \theta ij + \beta I(i = j)$, where $I(i = j) = 1$ if $i = j$; otherwise, it is equal to 0. This model improves the fit further ($L^2 = 6.36$ on 2 df , with a corresponding p value between .025 and .05), and the estimated association parameters are $\hat{\theta} = 0.47$ and $\hat{\beta} = 0.74$. Briefly, $\hat{\theta}$ accounts for a positive association between respondents' education attainment and oldest brothers' educational attainment (i.e., $\theta > 0$), and $\hat{\beta}$ accounts for still greater clustering of the two variables into the identical (i.e., diagonal) categories than under uniform association (i.e., $\beta > 0$).

The two refined models presented here only scratch the surface for illustrating the tools available for modeling association within the log-linear models framework. Indeed, the model of SYMMETRY (i.e., $\pi_{ij} = \pi_{ji}$), which corresponds to the log-linear model where $\lambda_i^R = \lambda_i^C$ and $\lambda_{ij}^{RC} = \lambda_{ji}^{RC}$, fits these data very well (i.e., $L^2 = 3.46$ on 3 df ; $p > .25$). The interested reader is referred to the references, as well as the references cited therein, to explore this subject further, including generalizations to nonlinear models for association in two-way tables and in multiway tables.

—Mark P. Becker

REFERENCES

- Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: John Wiley.
- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley.
- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: John Wiley.
- Mare, R. D. (1994). Discrete-time bivariate hazards with unobserved heterogeneity: A partially observed contingency table approach. In P. V. Marsden (Ed.), *Sociological methodology 1984* (pp. 341–383). Cambridge, MA: Blackwell.
- Powers, D. A., & Xie, Y. (2000). *Statistical methods for categorical data analysis*. San Diego: Academic Press.

LONGITUDINAL RESEARCH

Longitudinal research may best be defined by contrasting it with CROSS-SECTIONAL research. Cross-sectional research refers to research in which data are collected for a set of cases (individuals or aggregates such as cities or countries) on a set of variables (e.g., frequency of illegal behavior, attitudes toward globalization) and in which data collection occurs specifically (a) *at* a single time and (b) *for* a single time point or a single interval of time (hereafter, both will be referred to as *periods*). Analysis of purely cross-sectional data can examine *differences between cases* but not *changes within cases*. Different disciplines define *longitudinal research* differently, but a broad definition would include research in which (a) data are collected *for* more than one time period, (b) possibly but not necessarily involving collection of data *at* different time periods, and (c) permitting analysis that, at a minimum, involves the measurement and analysis of change over time within the same cases (individual or aggregate). Longitudinal data have been collected at the national level for more than 300 years, since 1665, and at the individual level since 1759, but longitudinal research in the social sciences has really flourished since the 1970s and 1980s (Menard, 2002).

Although some researchers would consider the collection of data *at* different time periods to be a defining characteristic of longitudinal research, it is also possible in principle to collect *retrospective* data, in which the data are collected *at* a single period but *for* more than one period. For example, a respondent may be asked to report on all of the times he or she has been a victim of crime, or has committed a crime, over a span of years up to and including her or his lifetime. In contrast to retrospective studies, *prospective* longitudinal data collection involves the repeated collection of data with usually short recall periods: crimes committed or victimizations experienced in the past week, month, or year or data on attitudes, beliefs, and sociodemographic characteristics at the instant the data are collected.

Longitudinal research serves two primary purposes: to describe patterns of change and to establish the direction (positive or negative, from *Y* to *X* or from *X* to *Y*) and magnitude (a relationship of magnitude zero indicating the absence of a relationship) of causal relationships. Change is typically measured with respect to one of two continua: chronological time (for

historical change) or age (for developmental change). Sometimes, it is difficult to disentangle the two. If older individuals are less criminal than younger individuals, is this because crime declines with age, or is it possible that older individuals were always less criminal (even when they were younger) and younger individuals will remain more criminal (even as they get older), or some combination of the two? With cross-sectional data, it is not possible to disentangle the effects of history and development. Even with longitudinal data, it may be difficult, but with data on the same individuals both at different ages and different time periods, it at least becomes possible.

There are several different types of longitudinal designs (Menard, 2002), as illustrated in Figure 1. In each part of Figure 1, the columns represent years, and the rows represent subjects, grouped by time of entry into the study. Thus, subjects enter the population or sample ("rows") at different times ("columns"), and subjects who have entered the study at different times may be in the study at the same time (more than one row outlined in the same column), except in the *repeated cross-sectional design*, in which different subjects are studied at each different time (no two rows outlined in the same column).

In a *total population design*, we attempt to collect data on the entire population at different time periods. An example of a total population design would be the Federal Bureau of Investigation's (annual) *Uniform Crime Report* data on arrests and crimes known to the police. Although coverage is not, in fact, 100% complete (the same is true for the decennial census), the intent is to include data on the entire U.S. population. From year to year, individuals enter the population (by birth) and exit the population (by death), so the individuals to whom the data refer overlap substantially but differ somewhat from one period to the next and may differ substantially over a long time span (e.g., from 1950 to the present). FBI data are used to measure aggregate rates of change or trends in arrests and crimes known to the police. The same design structure is present in the *Panel Study of Income Dynamics* (Hill, 1992), which is not a total population design but in which individuals may exit the sample by death or enter the sample by birth or marriage.

Repeated cross-sectional designs collect data on different samples in different periods. Because a new sample is drawn each year, there is in principle no overlap from one year to the next (although it is possible that independent samples will sample a few of

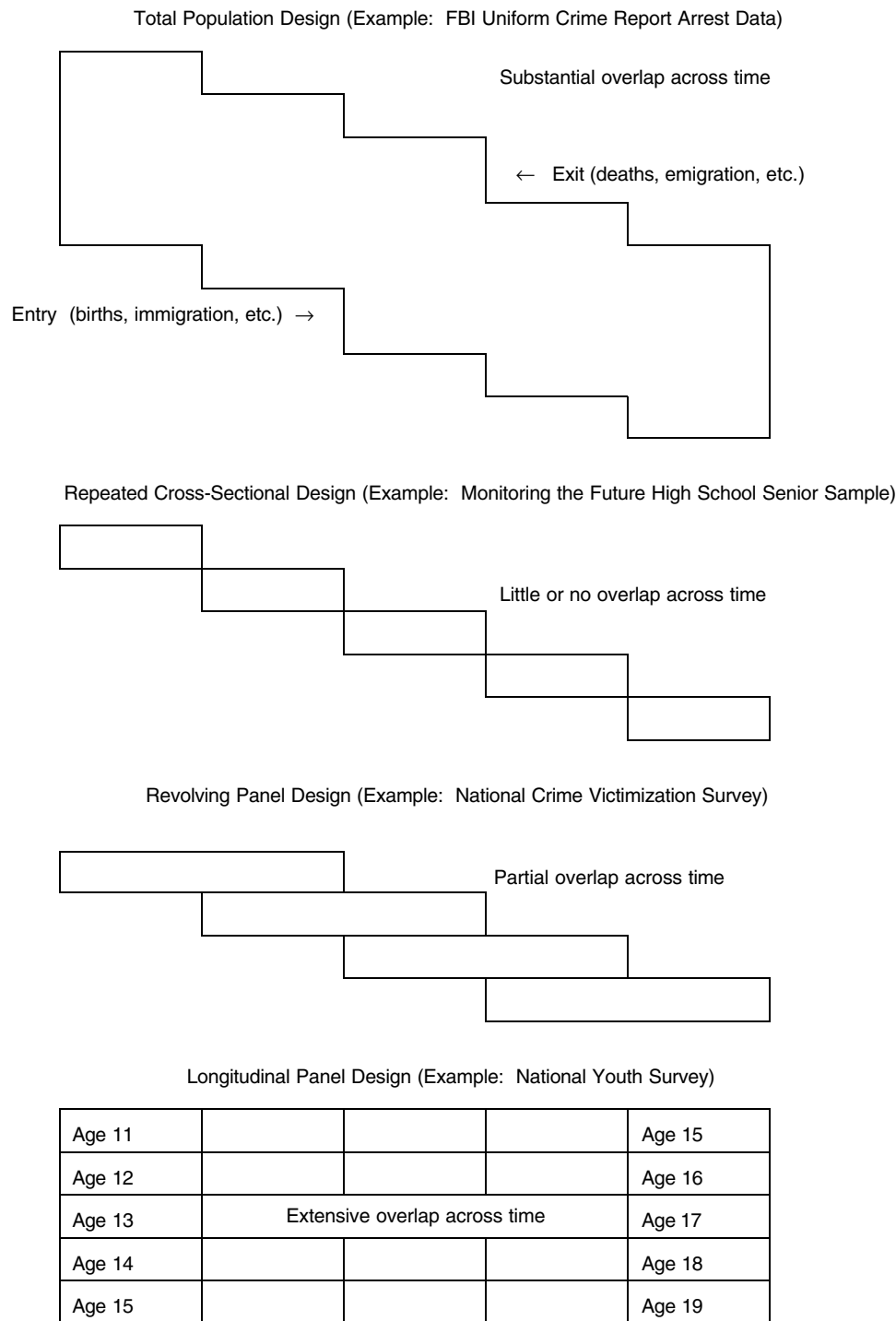


Figure 1 Types of Longitudinal Designs
SOURCE: Adapted from Menard (2002).

the same individuals). A good example of a repeated cross-sectional design is the General Social Survey, an annual general population survey conducted

by the National Opinion Research Center, which covers a wide range of topics (including marriage and family, sexual behavior and sex roles, labor

force participation, education, income, religion, politics, crime and violence, health, and personal happiness) and emphasizes exact REPLICATION of questions to permit comparisons across time to measure aggregate-level change (Davis & Smith, 1992).

Revolving panel designs are designs in which a set of respondents is selected, interviewed for more than one time period, then replaced by a new set of respondents. The revolving PANEL design is used in the National Crime Victimization Survey (NCVS; Bureau of Justice Statistics annual). Households are selected and interviewed seven times over a 3-year period: once at the beginning, once at the end, and at 6-month intervals between entry and exit. At the end of the 3-year period, a new household is selected to replace the old household. Replacement is staggered, so every 6 months, approximately one sixth of the households in the NCVS are replaced. It is important to note that the NCVS has historically been a sample of households, not individuals; whenever an individual or a family moved from a household, interviews were conducted not with the original respondents but with the (new) occupants of the original household. Because each household is replaced after 3 years, only short-term longitudinal data are available for individuals. At the national level, however, the NCVS is one of the most widely used sources of data on aggregate changes over time in rates of crime victimization and, by implication, the rate at which those offenses included in the NCVS (limited to offenses with identifiable victims) are committed.

The *longitudinal panel design* is the design most generally recognized in different disciplines as a true longitudinal design. An example of the longitudinal panel design is the National Youth Survey (NYS) (see, e.g., Menard & Elliott, 1990). The NYS first collected data on self-reported illegal behavior, including substance use, in 1977 but for the preceding year, 1976. It then collected data at 1-year intervals, for 1977 to 1980, and thereafter at 3-year intervals, for 1983 to 1992, and (as this is being written) additional data are being collected on the original respondents and their families in 2003. Unlike all of the types of longitudinal research discussed so far, there is no entry into the sample after the first year. The respondents interviewed in the most recent year were a subset of the respondents interviewed in the first year: Some of the original respondents were not interviewed because of death, refusal to participate, or failure to locate the

respondent. In contrast to the General Social Survey, FBI, and NCVS data, the NYS is used less to examine aggregate historical trends in crime than to examine intraindividual developmental trends, and causal relationships to test theories of crime. In this latter context, the NYS has been used to study the time ordering of the onset of different offenses and predictors of offending and to construct cross-time CAUSAL MODELS that can only be tested using longitudinal data (Menard & Elliott, 1990).

POTENTIAL PROBLEMS IN LONGITUDINAL RESEARCH

Longitudinal research potentially has all of the problems of cross-sectional research with respect to internal and external measurement validity; measurement RELIABILITY; SAMPLING ERROR; refusal to participate or NONRESPONSE to particular items; the appropriateness of questions to the population being studied; effects of interactions between subjects or respondents and interviewers, experimenters, or observers; relevance of the research; and research costs. Some of these issues are even more problematic for longitudinal research than for cross-sectional research. For example, biases in sampling may be amplified by repetition in repeated cross-sectional designs, and costs are typically higher for a multiple-year longitudinal study than for a single-year cross-sectional study. There are also additional dangers. As already noted, respondent recall failure may result in underreporting of illegal behavior if long recall periods are used. Respondents who are repeatedly interviewed may learn that giving certain answers results in follow-up questions and may deliberately or unconsciously avoid the burden imposed by those questions, a problem known as panel conditioning. Relatedly, the potential problem of interaction between the respondent or subject and an experimenter, interviewer, or observer producing invalid responses may be exacerbated when there is repeated contact between the experimenter, interviewer, or observer and the respondent or subject in a prospective longitudinal design. In later waves of a prospective panel study, respondents may have died or become incapable of participating because of age or illness or may refuse to continue their participation, or researchers may have difficulty locating respondents, resulting in panel attrition. In retrospective research, the corresponding problem is that individuals who should have been included to ensure a more REPRESENTATIVE SAMPLE may have

died or otherwise become unavailable before the study begins. Particularly in prospective longitudinal sample SURVEY research, an important question is whether attrition is so systematic or so great that the results of the study can no longer be generalized to the original population on which the sample was based.

Measurement used at the beginning of a longitudinal study may come to be regarded as obsolete later in the study, but changing the measurement instrument means that the data are no longer comparable from one time to the next. An example of this is the change in the way questions were asked in the NCVS in 1992. The new format produced a substantial increase in reported victimizations. A comparison was made between the rates of victimization reported using the old and the new method but only in a single year. Thus, it remains uncertain whether attempts to “adjust” the victimization rates to produce a closer correspondence between the old and the new methods are really successful, especially for examining long-term trends in victimization. In developmental research, a parallel problem is the inclusion of age-appropriate measures for the same concept (e.g., prosocial bonding) across the life course. For example, in adolescence, bonding may occur primarily in the contexts of family of orientation (parents and siblings) and school, but in adulthood, it may occur more in the contexts of family of procreation (spouse and children) and work. One is then faced with the dilemma of asking age-inappropriate questions or of using different measures whose comparability cannot be guaranteed for different stages of the life course.

In cross-sectional research, we may have MISSING DATA because an individual refuses to participate in the research (missing subjects) or because the individual chooses not to provide all of the data requested (missing values). In longitudinal research, we have the additional possibility that an individual may agree to participate in the research and provide the requested data at one or more periods but may refuse or may not be found at one or more other periods (missing occasions). Some techniques for analyzing longitudinal data are highly sensitive to patterns of missing data and cannot handle series of unequal lengths, thus requiring dropping all cases with missing data on even a single variable on just one occasion. Problems of missing data may be addressed in longitudinal research either by imputation of missing data (replacing the missing data by an educated guess, based on other data and the relationships among different variables in the study, sometimes including imputation of missing occasions)

or by using techniques that allow the use of partially missing data in the analysis (Bijleveld & van der Kamp, 1998, pp. 43–44).

LONGITUDINAL DATA ANALYSIS

Longitudinal research permits us to use dynamic models, models in which a change in one variable is explained as a result of change in one or more other variables. Although we often phrase statements about causal relationships as though we were analyzing change, with cross-sectional data, we are really examining how *differences* among individuals in one variable can be explained in terms of *differences* among those same individuals in one or more predictors. It has been relatively commonplace to draw conclusions about *intraindividual change*, or change within an individual respondent, from findings about *interindividual differences*, or differences between different respondents. Under certain conditions, it is possible to estimate dynamic models using cross-sectional data, but those conditions are restrictive and relatively rare in social science research (Schoenberg, 1977). It is generally more appropriate to use longitudinal data and to analyze the data using statistical techniques that take advantage of the longitudinal nature of the data. Longitudinal data analysis techniques, as described in Bijleveld and van der Kamp (1998), Menard (2002), and Taris (2000), include TIME-SERIES analysis for describing and analyzing change when there are a large number of time periods (e.g., changes in imprisonment rates as a function of changes in economic conditions), latent GROWTH CURVE MODELS and HIERARCHICAL LINEAR MODELS for describing and analyzing both short- and long-term change within individual subjects or cases and relationships among changes in different variables (e.g., changes in violent criminal offending as a function of changes in illicit drug use), and EVENT HISTORY ANALYSIS to analyze influences on the timing of qualitative changes (e.g., timing of the onset of illicit drug use as a function of academic performance).

Cross-sectional research cannot disentangle developmental and historical trends, and the description and analysis of historical change require the use of longitudinal data. The description and analysis of developmental trends can be attempted using cross-sectional data, but the results will not necessarily be consistent with results based on longitudinal data. Testing for the time ordering or sequencing of purported causes and effects in developmental data can

only be done with longitudinal data. Although there are cross-sectional methods for modeling patterns of mutual causation, more powerful models that allow the researcher to examine such relationships in more detail, including the explicit incorporation of the time ordering of cause and effect, require longitudinal data. Briefly, no analyses can be performed with cross-sectional data that cannot (by analyzing a single cross section) be performed with longitudinal data, but there *are* analyses that can be performed only with longitudinal, not cross-sectional, data. Cross-sectional data remain useful for describing conditions at a particular period, but increasingly in the social sciences, longitudinal data are recognized as best for research on causal relationships and patterns of historical and developmental change.

—Scott Menard

REFERENCES

- Bijleveld, C. C. J. H., & van der Kamp, L. J. Th. (with Mooijjaart, A., van der Kloot, W. A., van der Leeden, R., & van der Burg, E.). (1998). *Longitudinal data analysis: Designs, models, and methods*. London: Sage.
- Davis, J. A., & Smith, T. W. (1992). *The NORC General Social Survey: A user's guide*. Newbury Park, CA: Sage.
- Hill, M. S. (1992). *The panel study of income dynamics: A user's guide*. Newbury Park, CA: Sage.
- Menard, S. (2002). *Longitudinal research*. Thousand Oaks, CA: Sage.
- Menard, S., & Elliott, D. S. (1990). Longitudinal and cross-sectional data collection and analysis in the study of crime and delinquency. *Justice Quarterly*, 7, 11–55.
- Schoenberg, R. (1977). Dynamic models and cross-sectional data: The consequences of dynamic misspecification. *Social Science Research*, 6, 133–144.
- Taris, T. W. (2000). *A primer in longitudinal data analysis*. London: Sage.

LOWESS. See LOCAL REGRESSION

M

MACRO

Aggregate-level analysis is macro; individual-level analysis is micro. Thus, a political scientist may study elections at the macro level (e.g., presidential election outcomes for the nation as a whole) or at the micro level (e.g., presidential vote choice of the individual voter). Furthermore, an economist may study macroeconomic indicators, such as the country's gross domestic product, or the microeconomy of supply and demand for a particular product, say, automobiles. Finally, it should be noted that *macro* is the name applied to a computer program subroutine, as in SAS macros.

—Michael S. Lewis-Beck

MAIL QUESTIONNAIRE

Mailing QUESTIONNAIRES to PROBABILITY SAMPLES can be a cost-effective method to collect information when addresses are available for the POPULATION of interest (e.g., individuals or establishments). Single-mode surveys can be designed that employ mail questionnaires only, or these self-administered questionnaires may be part of a mixed-mode approach that includes other modes of administration (e.g., telephone, Web-based, or in-person INTERVIEWS). A frequently employed dual-mode design is to call sample members who have not returned mailed questionnaires to attempt interviews by telephone. Multimode designs can help reduce NONRESPONSE and often increase the representativeness of the resulting sample.

The SAMPLING FRAME for any RESEARCH DESIGN that includes a mail component can be a set of addresses, such as a list of employees of a firm or members of a professional organization. Note that the lists of addresses in telephone books cannot be relied on for sampling for population studies. The representativeness of samples drawn from telephone listings is compromised by unlisted numbers and households that either have no telephone service or rely solely on cell phones.

To identify members of rare populations for a mail SURVEY, one might use RANDOM-DIGIT DIALING (RDD) techniques. When a household is identified in which a member of the population of interest lives, the interviewer requests the address and a questionnaire can be mailed. Although the advantage of using RDD is that every household with a telephone has a known chance of selection, many people are unwilling to divulge their addresses to strangers who call on the telephone.

Respondent selection can be another challenge with mailed questionnaires (e.g., in situations when addresses, but not names, are known or when parents are required to randomly select one of their children to report on). There are techniques available for designating a respondent or subject (in the case of PROXY REPORTING), but these designs must be carefully planned, and the quality of the resultant sample depends on respondent compliance with sampling instructions.

Researchers have a toolbox of methods to use to develop survey instruments to collect data that can be used to adequately address their research questions. Instrument development often begins with a literature review to identify existing instruments that address

the issues at hand. It continues with FOCUS GROUPS conducted in the population(s) of interest to identify salient issues and the terminology that people use to talk about these topics. Candidate survey items are drafted about these domains of interest, and cognitive interviews, intensive one-on-ones with individuals from the target population, are conducted. The goal of these interviews is to identify survey questions and terminology that are not consistently understood. The instrument can then be revised and retested prior to fielding the survey. Instruments also often undergo expert review prior to launching the survey. Here, authorities in the field of interest are asked to propose domains and comment on the suitability of the planned instrument. The last step before fielding a mail questionnaire is a PRETEST. Individuals with characteristics that mirror the target population can be brought together for group pretests. First, participants complete the questionnaire, and then they are debriefed about their experiences with navigating the instrument and any difficulties they may have encountered in answering the questions. Group pretests also allow an estimation of how long it will take respondents to fill out the form.

Characteristics of the target population must be taken into consideration, both when deciding whether to use a mail questionnaire and in the design of mailed instruments. To minimize respondent burden, it is always important to consider the length and reading level of instruments. Instrument length and readability are especially key in mail surveys that include members of more vulnerable populations (e.g., those with poor reading or writing skills, people who are vision impaired, or those who may tire easily because of poor health).

With mailed questionnaires, there is no interviewer present to motivate respondents or to answer their questions. These instruments should be designed to make it easy for people to respond to questions and to successfully navigate skip patterns. CLOSED-ENDED questions that can be answered by simply checking a box or filling in a circle (for data entry using optical scanning techniques) are preferable to open-ended items, for which obtaining adequate answers and standardized CODING of responses can be problematic.

There are certain advantages to the use of mailed questionnaires, including better estimates of sensitive topics. People are more willing to report socially undesirable behaviors (e.g., substance abuse or violence) when there is no interviewer present. Another

advantage is the reduced cost of data collection relative to telephone and face-to-face administration. Mailing questionnaires can also be a successful approach to getting survey instruments past gatekeepers and into the hands of sample members, who may find the topic of study salient enough to warrant the investment of time necessary to respond. In addition, because respondents are not required to keep all response choices in memory while formulating their answers, mail administration allows longer lists of response choices than telephone.

The field period for a mail study is generally about eight weeks. Although there are a number of acceptable PROTOCOLS, standard mail survey research often includes three contacts by mail. First, a questionnaire packet is mailed that generally includes a cover letter outlining research goals and sponsors, the survey instrument, and a postage-paid envelope in which to return the completed questionnaire. This mailing is followed in about a week by a postcard reminder/thank you to all sample members. Around 2 to 3 weeks after the initial mailing, a replacement questionnaire packet is sent to all nonrespondents. Response rates are a function of the quality of the contact information available. Imprinting outgoing envelopes with a request for address corrections will yield updated addresses for some sample members who have moved.

The use of systematic quality control measures is key to successful data collection. Sample members should be assigned identification numbers to be affixed to questionnaire booklets to track response. Survey data are then entered, either manually or using optical scanning, to allow statistical analysis.

—Patricia M. Gallagher

REFERENCES

- Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method*. New York: John Wiley.
- Fowler, F. J., Jr. (2002). *Survey research methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Schwarz, N. A., & Sudman, S. (Eds.). (1996). *Answering questions*. San Francisco: Jossey-Bass.

MAIN EFFECT

The term *main effect* typically is used with reference to FACTORIAL DESIGNS in ANALYSIS OF VARIANCE. Consider a factorial design in which the levels of one factor, A, are crossed with the levels of another factor, B, as follows in Table 1.

Table 1

		Factor B			Marginal Mean
		Level B ₁	Level B ₂	Level B ₃	
Factor A	Level A1	M _{A1B1}	M _{A1B2}	M _{A1B3}	M _{A1}
	Level A2	M _{A2B1}	M _{A2B2}	M _{A2B3}	M _{A2}
	Marginal mean	M _{B1}	M _{B2}	M _{B3}	

The entries in each cell are means and the marginal entries are means for a given level of a factor collapsing across the levels of the other factor. For example, M_{A1} is the mean for Level 1 of Factor A collapsing across the levels of Factor B.

Informally, the term *main effect* refers to properties of the marginal means for a factor. Researchers use the term in different ways. Some will state that “there is a main effect of Factor A,” which implies that the marginal means for the factor are not all equal to one another in a POPULATION, a conclusion that is based on the results of a formal statistical test. Others simply refer to the “means comprising the main effect of Factor A,” drawing attention to the marginal means but without any implication as to their equality or the results of a statistical test that is applied to them.

There are different methods for calculating marginal means. One method uses unweighted means, whereby the marginal mean is the average of each cell mean that is being collapsed across. For example, M_{A1} is defined as (M_{A1B1} + M_{A1B2} + M_{A1B3})/3. A second method is based on weighted means, whereby each of the individual cell means that factor into the computation of the marginal mean is weighted in proportion to the SAMPLE size on which it is based. Means based on larger sample sizes have more weight in determining the marginal mean.

The term *main effect* also is used in LINEAR REGRESSION models. A “main effect” model is one that does not include any INTERACTION terms. For example, a “main effect” MULTIPLE REGRESSION model is

$$Y = a + b_1X + b_2Z + e,$$

whereas an “interaction model” is

$$Y = a + b_1X + b_2Z + b_3XZ + e.$$

The presence of the product term in the second equation makes this an interaction model. In regression contexts, a main effect refers to any continuous PREDICTOR in a REGRESSION model that is not part of a product

term or an interaction term. A main effect also refers to any dummy variable or set of DUMMY VARIABLES that is not part of a product term or interaction term. In the following model of all continuous predictors,

$$Y = a + b_1X + b_2Z + b_3XZ + b_4Q + e.$$

The coefficient for Q might be referred to as representing a “main effect” of Q , but the coefficients for X , Z , and XZ would not be so characterized.

—James J. Jaccard

MANAGEMENT RESEARCH

The academic discipline of management evolved rapidly in the 1960s within the United States and in the 1980s in most other parts of the world and is consequently a relative newcomer to the pantheon of the social sciences. In the early days, it drew very heavily on related disciplines such as economics, sociology, anthropology, statistics, and mathematics. But more recently, there has been a growing appreciation that these related disciplines are inadequate in themselves for handling the complex theoretical and practical problems posed in the management domain.

Three distinctive features of management research imply a need both to adapt traditional social science methods and to identify new ways of conducting research. First, there is a marked difference between the practice of management and the academic theories about management. The former draws eclectically on ideas and methods to tackle specific problems; the latter adopts distinct disciplinary frameworks to examine aspects of management. The adoption of disciplines, ranging from highly quantitative subjects such as finance, economics, and operations research to highly qualitative subjects such as organization behavior and human resource development, makes it hard to achieve consensus on what constitutes high-quality research. It has also led to a sustained debate about the relative merits of adopting monodisciplinary or transdisciplinary perspectives (Kelemen & Bansal, 2002; Tranfield & Starkey, 1998).

The second issue concerns the stakeholders of management research because managers, the presumed objects of study, will normally have the power to control ACCESS and may also be in a position to provide or withdraw financial sponsorship from the researcher.

This means that management research can rarely divorce itself from political and ethical considerations; it also means that specific methods have to be adapted to meet the constraints of managerial agendas (Easterby-Smith, Thorpe, & Lowe, 2002). Thus, for example, it is rarely possible to adopt the emergent or THEORETICAL SAMPLING strategies recommended by the proponents of GROUNDED THEORY because managers may insist on prior negotiation of interview schedules and will severely restrict access due to the opportunity costs of having researchers taking up the time of employees. Hence, the canons of grounded theory have to be adapted in most cases (Locke, 2000).

Third, the practice of management requires both thought and action. Not only do most managers feel that research should lead to practical consequence, but they are also quite capable of taking action themselves in the light of research results. Thus, research methods either need to incorporate with them the potential for taking actions, or they need to take account of the practical consequences that may ensue with or without the guidance of the researcher. At the extreme, this has led to a separation between pure researchers who try to remain detached from their objects of study and ACTION RESEARCHERS or consultants who create change to learn from the experiences. This has also led to the development of a range of methods such as participative inquiry (Reason & Bradbury, 2000) and Mode 2 research (Huff, 2000), which have considerable potential for research into management.

Of course, these three features are widely encountered across the social sciences, but it is their simultaneous combination that is driving the methodological adaptations and creativity encountered in some management research. In some cases, this has resulted in the adaptation and stretching of existing methods; in other cases, it is leading to the evolution of new methods and procedures. Furthermore, these evolving methods are likely to have a greater impact on the social sciences in general due to the sheer scale of the academic community in the field of management and business. At the time of this writing, roughly one third of all social scientists in the United Kingdom work within business schools. No doubt things have progressed even further in the United States.

—Mark Easterby-Smith

REFERENCES

Easterby-Smith, M., Thorpe, R., & Lowe, A. (2002). *Management research: An introduction* (2nd ed.). London: Sage.

Huff, A. S. (2000). Changes in organizational knowledge production. *Academy of Management Review*, 25(2), 288–293.

Keleman, M., & Bansal, P. (2002). The conventions of management research and their relevance to management practice. *British Journal of Management*, 13, 97–108.

Locke, K. D. (2000). *Using grounded theory in management research*. London: Sage.

Reason, P., & Bradbury, H. (2000). *Handbook of action research: Participative inquiry and practice*. London: Sage.

Tranfield, D., & Starkey, K. (1998). The nature, organization and promotion of management research: Towards policy. *British Journal of Management*, 9, 341–353.

MANCOVA. See MULTIVARIATE ANALYSIS OF COVARIANCE

MANN-WHITNEY U TEST

This is a frequently used test to determine whether two independent groups belong to the same population. It is considered to be one of the most powerful NON-PARAMETRIC TESTS available and it is appropriate for data that is at least ordinal. Two groups developed what is essentially the same test, the *Mann-Whitney U-test* and the *Wilcoxon rank-sum test*, thus references in the literature tend to refer to them as one test, the Wilcoxon-Mann-Whitney test. The actual calculations may vary from one reference to another, but the same end is achieved. It is also of use as an alternative to the *T-TEST* when its assumptions of normality and homogeneity of variance cannot be met (Siegel and Castellan, 1988). When the sample sizes are small, it is necessary to have special tables, or depend on a computer-based statistical package to tell what the probability is.

The test is based upon the rank order of the data and is asking whether the two underlying population distributions the same. This is resolved by determining whether the sum of the ranks of one sample is sufficiently different from the overall mean of the ranks of both of the groups to indicate that it is not part of a common POPULATION. The SAMPLING DISTRIBUTION again says that any sample from a population will have a sum of ranks close to the population mean but not necessarily identical. The distribution of the sum of ranks for “large” samples is assumed to be normal with a population MEAN defined as

$$\mu = 1/2n_1(n_1 + n_2 + 1)$$

and a standard error of the mean found by

$$SEM = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}.$$

Some texts suggest that large can be 10 or more in each group; others will maintain a minimum of 25. In any case, the following approximation is more accurate the larger the sample. It is not difficult to see that the test is based upon the z-TEST,

$$z = \frac{\bar{x} - \mu}{SEM},$$

which, when using the sum of the ranks, W_1 , for one of the groups as the sample mean, becomes, using the Wilcoxon approach,

$$z = \frac{W_1 - 1/2 n_1 (n_1 + n_2 + 1)}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

where

- W_1 = the sum of the ranks for one of the two groups,
- n_1 = the sample size of one group (≥ 10),
- n_2 = the sample size of the other group (≥ 10).

Thus, to carry out the test on sample data, the first task is to arrange all the raw data in rank order and use the order as the data. This will require that when “ties” in scores occur, the subjects having identical scores be assigned equivalent ranks. The z-score is then checked against the appropriate point in the table for the choice of α .

For example, the data in Table 1 shows scores on a Spanish language vocabulary test for two groups after two weeks, each learning the same material by different approaches. The groups were formed by randomly assigning students from a second year class. Group A learned by teacher lead discussions in class with visual aids, and Group B learned by computer based exercises, working on their own. In this case, the score distributions were highly skewed, thus the second column for each set of raw data reflects the ranks for the whole set (both groups combined).

The question is, do they still belong to the same population, allowing for natural variation? The ranks are then added for each group and one of them becomes W_1 , as shown. The calculation of z then becomes

$$z = \frac{W_1 - 1/2 n_1 (n_1 + n_2 + 1)}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

Table 1 Data From a Spanish-Language Test for Two Groups, With Columns Added to List Rank Order

	Data A	Rank A	Data B	Rank B
	86	2	92	1
	78	5	85	3
	75	8	80	4
	73	10.5	77	6
	73	10.5	76	7
	72	12	74	9
	70	14	71	13
	68	16	69	15
	64	17.5	64	17.5
	60	20	61	19
	55	22	57	21
	50	25	52	23
	46	26	51	24
$W =$		188.5		162.5
$n =$		13		13

NOTE: Tied ranks are the average of the sequential ranks.

$$z = \frac{188.5 - 1/2(13)(13 + 13 + 1)}{\sqrt{\frac{(13)(13)(13+13+1)}{12}}}$$

$$z = \frac{188.5 - 175.5}{\sqrt{380.25}} = \frac{13}{19.5} = 0.67$$

If α were chosen to be 0.05, the minimum value for z would be 1.96 for a significant difference. Therefore, this result indicates that the two groups are not significantly different in their distributions of ranks and would still seem to belong to the same population. In other words, at least for this exercise, there was no difference due to learning approach, assuming all other possible extraneous variables were controlled.

The equivalent Mann-Whitney approach calculates U_1 , which is

$$U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - W_1,$$

and then the calculation of z for U_1 (producing the same value as above) becomes

$$z = \frac{U_1 - (n_1 n_2 / 2)}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}.$$

As Howell (2002) notes, the simple linear relationship results in the two differing only by a constant for any set of samples, making it easy to convert one to the other if the need arises.

—Thomas R. Black

REFERENCES

- Black, T. R. (1999). *Doing Quantitative Research in the Social Sciences: An Integrated Approach to Research Design, Measurement, and Statistics*. London: Sage Publications.
- Grimm, L. G. (1993). *Statistical Applications for the Behavioral Sciences*. New York: John Wiley & Sons.
- Howell, D. C. (2002). *Statistical Methods for Psychology*. (5th ed) Pacific Grove, CA: Duxbury.
- Siegel, S., and Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

MANOVA. See MULTIVARIATE ANALYSIS OF VARIANCE

MARGINAL EFFECTS

In REGRESSION analysis, data analysts are oftentimes interested in interpreting and measuring the effects of INDEPENDENT (or explanatory) VARIABLES on the DEPENDENT (or response) VARIABLE. One way to measure the effects of independent variables is to compute their *marginal effects*. The marginal effect of an independent variable measures the impact of change in an independent variable (e.g., X_i) on the expected change in the dependent variable (e.g., Y) in a regression model, especially when the change in the independent variable is infinitely small or merely marginal.

The marginal effect of an independent variable X on the dependent variable Y can be computed by taking the partial derivative of $E(Y|X)$ with respect to X if the independent variable is *continuous* and thus differentiable. Consider a linear regression model with two independent variables,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

where Y_i is the value of the dependent variable for the i th observation, X_{1i} is the value of the continuous independent variable, X_{2i} is the value of a dichotomous variable that equals either 0 or 1, β s are regression coefficients, and ε_i is the error term, which is assumed to be distributed independently with zero mean and constant variance σ^2 . Given the assumption of the classical regression model, the conditional mean of the dependent variable Y is simply $E(Y|X) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$. Thus, the marginal effect of X_1 on the expected change in Y in this linear regression model

can be computed by obtaining the partial derivative with respect to X_1 , that is,

$$\frac{\partial E(Y|X, \beta)}{\partial X_1} = \frac{\partial(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{\partial X_1} = \beta_1. \quad (1)$$

The marginal effect of X_2 , however, cannot be computed by taking the partial derivative because X_2 is a dichotomous variable and not differentiable. Greene (2002) and Long (1997) suggest using the *discrete change* in $E(Y|X, \beta)$ as the dichotomous variable (in this case, X_2) changes from 0 to 1, holding all other variables constant (usually at their mean) to interpret the marginal effect of a dichotomous independent variable. Symbolically, that is,

$$\begin{aligned} \frac{\Delta E(Y|X, \beta)}{\Delta X_2} &= (\beta_0 + \beta_1 X_1 + \beta_2 \times 1) \\ &\quad - (\beta_0 + \beta_1 X_1 + \beta_2 \times 0) = \beta_2. \quad (2) \end{aligned}$$

From equations (1) and (2), we know that the marginal effect of an independent variable in a linear regression model is simply equal to the parameter β . Thus, interpretation of parameter estimates ($\hat{\beta}_1$ and $\hat{\beta}_2$, in this case) might be described as follows: Holding all other variables constant, a unit increase in X_1 led to, on average, about a $\hat{\beta}$ increase (or decrease, depending on the sign of $\hat{\beta}$) in the dependent variable Y . Note that a unit change for a dichotomous variable is equal to the variable's range because it only takes on value of 0 or 1. Thus, interpretation of regression on DUMMY VARIABLES might be described as follows: Holding all other variables constant, having the attribute of X_1 (as opposed to not having the attribute) led to, on average, an increase (or decrease) of $\hat{\beta}$ in Y .

Interpretation of marginal effects of independent variables in NONLINEAR regression models is far more complicated because the effect of a unit change in an independent variable depends on the values of *all* independent variables and *all* of the model parameters. As a result, in almost no case is the reported coefficient equal to a marginal effect in nonlinear models such as logit, probit, multinomial logit, ordered probit and logit, and event count models.

Consider a binary choice model such as a logit or a probit model,

$$\Pr(Y = 1|X) = F(X\beta),$$

in which the link function F is the cumulative distribution for the NORMAL DISTRIBUTION or for the logistic

distribution. In either case, the marginal effect of an independent variable is the expected change in the probability due to change in that independent variable. This can also be computed by taking the partial derivative if the data are differentiable. For example, the marginal effect of the k th independent variable (X_k) is given by

$$\frac{\partial \Pr(Y = 1 | \mathbf{X}, \boldsymbol{\beta})}{\partial X_k} = \frac{\partial F(\mathbf{X}\boldsymbol{\beta})}{\partial X_k} = f(\mathbf{X}\boldsymbol{\beta})\beta_k, \quad (3)$$

where f is either the probability density function for the standard normal distribution or for the logistic distribution, $\mathbf{X}$ is the matrix of values of independent variables for all observations, and $\boldsymbol{\beta}$ is the vector of parameters. Because $f(\mathbf{X}\boldsymbol{\beta})$, which is also known as the scale factor, is nonnegative by definition, equation (3) shows that the direction (or sign) of the marginal effect of X_k is completely determined by β_k . However, the magnitude of its marginal effect is jointly determined by both the values of β_k and $f(\mathbf{X}\boldsymbol{\beta})$. Thus, interpretation of a marginal effect in nonlinear models needs to further consider the nature of the dependent variable (e.g., continuous, dichotomous, ordinal, or nominal), the problem being analyzed, and the probability distribution chosen by a researcher. For more information on the issue regarding marginal effects, one can refer to Long (1997), Greene (2002), and Sobel (1995).

—Cheng-Lung Wang

REFERENCES

- Greene, W. H. (2002). *Econometric analysis* (5th ed.). New York: Prentice Hall.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 1–38). New York: Plenum.

MARGINAL HOMOGENEITY

Marginal homogeneity is a class of models that equates some of the marginal distributions of a multi-dimensional contingency table. For a $R \times R$ table, univariate marginal homogeneity is satisfied if the cell probabilities π_{ij} satisfy

$$\sum_j \pi_{ij} = \sum_j \pi_{ji} \quad \text{for all } i.$$

For a $R \times R \times R$ table, bivariate marginal homogeneity is satisfied if the cell probabilities π_{ijk} satisfy

$$\sum_k \pi_{ijk} = \sum_k \pi_{ikj} = \sum_k \pi_{kij} \quad \text{for all pairs } (i, j).$$

—Marcel A. Croon

MARGINAL MODEL

Statistical models that impose constraints on some (or all) marginal distributions of a set of variables as well as, eventually, on their joint distribution are called marginal models. In a multinormal distribution, modeling the marginal distributions does not create special problems because the first- and second-order moments of any marginal distribution are equal to their counterparts in the joint distribution. For CATEGORICAL variables, such a one-to-one correspondence between the parameters of the joint and the marginal distributions does not exist in general, and one has to resort to specially developed procedures for analyzing data by means of marginal models.

Table 1 contains data for a response variable Y that is measured at three time points.

Table 1 Observed Frequency Distribution

	$Y_3 = 1$		$Y_3 = 2$	
	$Y_2 = 1$	$Y_2 = 2$	$Y_2 = 1$	$Y_2 = 2$
$Y_1 = 1$	121	11	41	44
$Y_1 = 2$	4	2	1	12

SOURCE: Data from the National Youth Survey from the Inter-University Consortium for Political and Social Research (ICPS), University of Michigan.

The data reported here were collected on 236 boys and girls who were repeatedly interviewed on their marijuana use. Here we use the data collected in 1976 (Y_1), 1978 (Y_2), and 1980 (Y_3) with responses coded as 1 = never used marijuana before and 2 = used marijuana before.

The corresponding joint probability distribution of Y_1 , Y_2 , and Y_3 will be represented by $\pi_{123}(y_1, y_2, y_3)$; the bivariate marginal distribution of Y_1 and Y_2 is then given by

$$\pi_{12}(y_1, y_2) = \sum_{y_3} \pi_{123}(y_1, y_2, y_3)$$

and the univariate marginal distribution of Y_1 by

$$\pi_1(y_1) = \sum_{y_2} \sum_{y_3} \pi_{123}(y_1, y_2, y_3).$$

Other marginal distributions can be defined similarly.

The hypothesis that Y_3 and Y_1 are independent given Y_2 imposes constraints on the joint distribution of the three variables, and these can be tested by fitting the LOG-LINEAR MODEL $[Y_1 Y_2, Y_1 Y_3]$ in the original three-dimensional contingency table. If a model only imposes constraints on the joint distribution, it is not a marginal model.

However, instead of formulating hypotheses about the structure of the entire table, one could be interested in hypotheses that state that some features of the marginal distributions of the response variable do not change over time. The hypothesis of univariate marginal homogeneity states that the univariate marginal distribution of Y is the same at the three time points:

$$\pi_1(y) = \pi_2(y) = \pi_3(y) \quad \text{for all } y.$$

In our example, the corresponding observed univariate marginal response distributions are presented in Table 2.

Table 2 Univariate Marginals

	Y_1	Y_2	Y_3
1	217	167	138
2	19	69	98
	236	236	236

For these data, the hypothesis that all three distributions arise from the same probability distribution has to be rejected with $L^2 = 82.0425$ and $df = 2$ ($p < .001$). Hence, in this panel study, there is clear evidence for a significant increase in marijuana use over time. Note that this hypothesis cannot be tested by the conventional CHI-SQUARE TEST for independence because the three distributions are not based on independent samples.

Alternatively, one could be interested in the hypothesis that some of the bivariate marginal distributions are identical, such as

$$\pi_{12}(y_1, y_2) = \pi_{23}(y_1, y_2) \\ \text{for all pairs of scores } (y_1, y_2).$$

The corresponding observed bivariate distributions for (Y_1, Y_2) and (Y_2, Y_3) are, respectively, given by Tables 3 and 4.

The hypothesis that both bivariate marginal distributions derive from the same bivariate probability

Table 3 Bivariate Marginal for (Y_1, Y_2)

	$Y_2 = 1$	$Y_2 = 2$
$Y_1 = 1$	162	55
$Y_1 = 2$	5	14

Table 4 Bivariate Marginal for (Y_2, Y_3)

	$Y_3 = 1$	$Y_3 = 2$
$Y_2 = 1$	125	42
$Y_2 = 2$	13	56

distribution is also rejected because $L^2 = 86.0024$ with $df = 4$ ($p < .001$).

A less restrictive hypothesis would equate corresponding association parameters from each bivariate distribution. The assumption of equality of corresponding local ODDS RATIOS in the bivariate marginals π_{12} and π_{23} leads to constraints of the following type:

$$\frac{\pi_{12}(y_1, y_2)\pi_{12}(y_1 + 1, y_2 + 1)}{\pi_{12}(y_1 + 1, y_2)\pi_{12}(y_1, y_2 + 1)} \\ = \frac{\pi_{23}(y_1, y_2)\pi_{23}(y_1 + 1, y_2 + 1)}{\pi_{23}(y_1 + 1, y_2)\pi_{23}(y_1, y_2 + 1)} \\ \text{for all pairs of scores } (y_1, y_2).$$

In a 2×2 table, there is only one odds ratio to consider. For the present data, it is equal to 8.247 in the marginal (Y_1, Y_2) table and to 12.8205 in the marginal (Y_2, Y_3) table. The hypothesis that both odds ratios are equal cannot be rejected now because $L^2 = 0.4561$ with $df = 1$ ($p = .499$). The estimate of the constant odds ratio in both marginal tables is 11.3463. This result indicates that a strong positive and constant association exists between responses that are contiguous in time. For variables with more than two response categories, alternative marginal models can be formulated that equate corresponding adjacent, cumulative, or global odds ratios.

Finally, one can also impose constraints on conditional distributions defined in terms of the marginals. For example, the hypothesis that the conditional distribution of Y_2 given Y_1 is equal to the conditional distribution of Y_3 given Y_2 leads to the following constraints:

$$\frac{\pi_{12}(y_1, y_2)}{\pi_1(y_1)} = \frac{\pi_{23}(y_1, y_2)}{\pi_2(y_1)} \\ \text{for all pairs of scores } (y_1, y_2).$$

For the present data, we have

$$\begin{aligned}\Pr(Y_2 = 2|Y_1 = 1) &= 0.2535, \\ \Pr(Y_2 = 2|Y_1 = 2) &= 0.7368, \\ \Pr(Y_3 = 2|Y_2 = 1) &= 0.2515, \text{ and} \\ \Pr(Y_3 = 2|Y_2 = 2) &= 0.8116.\end{aligned}$$

The hypothesis that corresponding TRANSITION RATES are constant over time could not be rejected with $L^2 = 0.5016$ for $df = 2$ ($p = .778$). The estimates of these constant transition probabilities were the following:

$$\begin{aligned}\Pr(Y_{t+1} = 2|Y_t = 1) &= 0.2529, \text{ and} \\ \Pr(Y_{t+1} = 2|Y_t = 2) &= 0.7959.\end{aligned}$$

These results indicate that about 80% of the users at a particular time point continue to do so at the next time point, whereas in the same period, about 25% of the nonusers become users.

Estimating the parameters under a marginal model and testing whether the model provides an acceptable fit to the data both require specifically developed methods. In many cases, the ensuing marginal model can be formulated as a generalized log-linear model for which appropriate estimation and testing procedures have been developed. These procedures perform MAXIMUM LIKELIHOOD ESTIMATION of the probabilities in the joint distribution under the constraints imposed on the joint and the marginal distributions and allow conditional and unconditional tests for the assumed model.

—Marcel A. Croon

REFERENCES

- Bergsma, W. (1997). *Marginal models for categorical data*. Tilburg, The Netherlands: Tilburg University Press.
- Croon, M. A., Bergsma, W., & Hagenaars, J. A. (2000). Analyzing change in categorical variables by generalized log-linear models. *Sociological Methods & Research*, 29, 195–229.
- Vermunt, J. K., Rodrigo, M. F., & Ato-Garcia, M. (2001). Modeling joint and marginal distributions in the analysis of categorical panel data. *Sociological Methods & Research*, 30, 170–196.

MARGINALS

Marginals or marginal distributions are obtained by collapsing a multivariate joint distribution over one or more of its variables. If X , Y , and Z are three

discrete random variables with probability distribution $\pi_{XYZ}(x, y, z)$, the marginal distribution of X and Y is given by

$$\pi_{XY}(x, y) = \sum_z \pi_{XYZ}(x, y, z).$$

The marginal distribution of X is given by

$$\pi_X(x) = \sum_y \sum_z \pi_{XYZ}(x, y, z).$$

For continuous variables, summing over a variable is replaced by integrating over it.

—Marcel A. Croon

MARKOV CHAIN

A Markov chain describes one's probability to change state. The focus is on the probability to change state in a discrete period of time, for instance, from political party A to another political preference. It is a process without memory: The probability to be in a state at the present measurement occasion only depends on one's state at the last previous occasion. States at earlier occasions have no additional predictive value. Once we observe a cross-table of two subsequent occasions, we can estimate the transition probability for all combinations in the state space: the probability of A given state A, $\tau_{A|A}$; the probability of other, O, given A, $\tau_{O|A}$; and so on, $\tau_{A|O}$ and $\tau_{O|O}$. Generalization to more states is straightforward.

The Markov chain is due to the Russian mathematician Andrei A. Markov. Interestingly, the Markov chain allows prediction of the distribution over states in the future. With the data example of Table 1, it is easily verified that Party A, starting with 67%, has lost 5% support after 1 year. In the end, this Markov chain will reach an equilibrium, with party A having the preference of 58% of the population, irrespective of the original distribution over states. This can be computed by matrix multiplication of the matrix with transition probabilities, T , with T itself and the result again with T and so on, until all rows of the resulting matrix are the same.

The fit of a Markov chain can be tested if it is applied to PANEL data on three or more occasions. When doing so, the Markov chain predicts too much long-term change in most social science applications. Blumen, Kogan, and McCarthy (1968) therefore postulated a mixture of two sorts of people: Some people change

Table 1 Change of State in Political Preferences

Original State	One Year Later; Frequencies		One Year Later; Transition Probabilities		7-Year Transition Probabilities Assuming a Markov Chain	
	Party A	Other	Party A	Other	Party A	Other
Party A	750	250	0.75	0.25	0.58	0.42
Other	175	325	0.35	0.65	0.58	0.42

state, and others do not. With a panel observed at three or more occasions, the proportions of “movers” and “stayers” can be estimated.

A pitfall in the analysis of change is that measurement error can blur the outcomes. If independent between occasions, such error will inflate the observed change, unless it is incorporated in a model. The LATENT MARKOV MODEL, also known as the hidden Markov model, allows erroneous responses. Each answer in the state space can occur, but its probability depends on the state one is in. People may have a high probability to answer correctly, and lower probabilities may hold for erroneous responses. This model can be estimated from a panel observed at three or more occasions.

Both approaches can be combined, for instance, assuming that part of the population behaves according to a latent Markov chain and another part consists of stayers. For accurate estimation of such a mixed, partially latent Markov model, it is better to have multiple indicators at each occasion.

A complication arises if people tend to return to states they were previously in, for instance, unemployment in labor market status research. Then previous state membership can be modeled as an exogenous variable, influencing the probability to become unemployed.

Finally, transition probabilities in discrete time are in fact the result of more basic flow parameters in continuous time—TRANSITION RATES or HAZARD RATES—that apply in an infinitesimally small period of time.

—Frank van de Pol and Rolf Langeheine

REFERENCES

Blumen, I., Kogan, M., & McCarthy, P. J. (1968). Probability models for mobility. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in mathematical social science* (pp. 318–334). Cambridge: MIT Press.

Langeheine, R., & van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods & Research*, 18, 416–441.

Vermunt, J. K., Rodrigo, M. F., & Ato-Garcia, M. (2001). Modeling joint and marginal distributions in the analysis of categorical panel data. *Sociological Methods & Research*, 30, 170–196.

MARKOV CHAIN MONTE CARLO METHODS

Markov chain Monte Carlo (MCMC) methods are SIMULATION techniques that can be used to obtain a random sample approximately from a probability distribution *p* by constructing a MARKOV CHAIN whose stationary distribution is *p* and running the Markov chain for a large number of iterations. One of the most popular application areas of MCMC methods is BAYESIAN INFERENCE. Here MCMC methods make it possible to numerically calculate summaries of Bayesian POSTERIOR DISTRIBUTIONS that would not be possible to calculate analytically.

In many applications, researchers find it useful to be able to sample from an arbitrary probability distribution *p*. This is particularly the case when one wishes to calculate a numerical approximation to some integral, as is commonly the case in Bayesian inference and, to a lesser extent, likelihood inference. In situations when *p* is a member of a known family of distributions, it is sometimes possible to use standard pseudo-random number generation techniques to produce an independent stream of pseudo-random deviates directly from *p*. However, in many practical applications, *p* is not a member of a known family of distributions and cannot be sampled from directly using standard methods. Nonetheless, MCMC methods can be used in such situations and, in general, are relatively easy to implement.

MCMC methods work not by generating an independent sample directly from *p* but rather by constructing a sequence of dependent draws from a particular stochastic process whose long-run equilibrium distribution is *p*. By constructing a long enough sequence of such draws, we can think of the later draws

as being approximately from p . More specifically, MCMC methods provide recipes for constructing Markov chains whose equilibrium (or stationary) distribution is the distribution p of interest.

A number of complications arise because the sequence of MCMC draws is not an independent sequence of draws taken directly from p . One complication concerns the number of initial so-called *burn-in* iterations to discard. The idea here is that because the MCMC draws are only approximately from p after running the Markov chain for a long period of time, the initial MCMC draws are not likely to be representative of draws from p and so should be discarded. Another complication involves the determination of how long the Markov chain should be run to obtain a sample that is sufficiently close to a direct sample from p for the application at hand. These issues are discussed at some length in standard reference texts such as Robert and Casella (1999).

Two basic types of MCMC algorithms are commonly used by researchers (although it should be noted that one is actually a special case of the other). The first is what is known as the Metropolis-Hastings algorithm. The second is a type of Metropolis-Hastings algorithm referred to as the GIBBS SAMPLING algorithm.

THE METROPOLIS-HASTINGS ALGORITHM

The Metropolis-Hastings algorithm is a very general MCMC algorithm that can be used to sample from probability distributions that are only known up to a normalizing constant. The Metropolis-Hastings algorithm works by generating candidate values from a candidate-generating density that is easily simulated from and then probabilistically accepts or rejects these candidate values in such a way as to produce a sample approximately from the distribution p of interest.

Suppose we want to draw a random sample from $p(\theta)$ where θ may be multidimensional. The Metropolis-Hastings algorithm is shown in Figure 1.

```

initialize  $\theta^{(0)}$ 
for (i in 1 to m){
  generate  $\theta_{can}^{(i)}$  from  $q(\theta|\theta^{(i-1)})$ 
  calculate  $\alpha \leftarrow \frac{f(\theta_{can}^{(i)}) q(\theta^{(i-1)}|\theta_{can}^{(i)})}{f(\theta^{(i-1)}) q(\theta_{can}^{(i)}|\theta^{(i-1)})}$ 
  set  $\theta^{(i)} \leftarrow \begin{cases} \theta_{can}^{(i)} & \text{with probability } \min\{1, \alpha\} \\ \theta^{(i-1)} & \text{with probability } 1 - \min\{1, \alpha\} \end{cases}$ 
  store  $\theta^{(i)}$ 
}
```

Figure 1

where $f(\theta)$ is some function proportional to $p(\theta)$, and $q(\theta|\theta^{(i-1)})$ is the candidate-generating density. Note that the candidate-generating density depends on the current value of $\theta^{(i-1)}$. In situations when the candidate-generating density is symmetric—that is, $q(\theta|\theta^{(i-1)})$ is equal to $q(\theta^{(i-1)}|\theta)$ —then these terms cancel in the formula for the acceptance probability α .

THE GIBBS SAMPLING ALGORITHM

The Gibbs sampling algorithm is a special case of the Metropolis-Hastings algorithm in which the acceptance probability is always equal to 1. The Gibbs sampling algorithm generates a sample approximately from some joint distribution p by iteratively sampling from the conditional distributions that make up p .

To see how the Gibbs sampling algorithm works, suppose we have a joint distribution $p(\theta_1, \dots, \theta_k)$, where it is possible to break the arguments of p into k distinct blocks. Let $p_1(\theta_1|\theta_2, \dots, \theta_k)$ denote the conditional distribution of θ_1 given the remaining $k-1$ blocks, with similar notation for the other $k-1$ full conditional distributions. The Gibbs sampling algorithm works as shown in Figure 2.

```

initialize  $\theta_1^{(0)}, \dots, \theta_k^{(0)}$ 
for (i in 1 to m){
  generate  $\theta_1^{(i)}$  from  $p_1(\theta_1|\theta_2^{(i-1)}, \dots, \theta_k^{(i-1)})$ 
  generate  $\theta_2^{(i)}$  from  $p_2(\theta_2|\theta_1^{(i)}, \theta_3^{(i-1)}, \dots, \theta_k^{(i-1)})$ 
  .
  .
  .
  generate  $\theta_k^{(i)}$  from  $p_k(\theta_k|\theta_1^{(i)}, \dots, \theta_{k-1}^{(i)})$ 
  store  $\theta_1^{(i)}, \dots, \theta_k^{(i)}$ 
}
```

Figure 2

HISTORICAL DEVELOPMENT

MCMC methods have their origins in statistical physics. Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) proposed what is now known as the Metropolis algorithm to perform the numerical integrations needed to explore certain properties of physical systems. The Metropolis algorithm was generalized by Hastings (1970) to work with nonsymmetric

candidate-generating densities. Geman and Geman (1984) introduced what is now referred to as the Gibbs sampling algorithm within the context of image restoration problems. Despite these (and several other) earlier works, statisticians did not fully embrace the potential of MCMC methods until the work of Gelfand and Smith (1990), which showed how MCMC techniques could be used to perform Bayesian inference for models that, up to that point, were regarded as computationally infeasible.

EXAMPLE: SAMPLING FROM A BAYESIAN POSTERIOR DISTRIBUTION

Consider the Bayesian treatment of the simple LOGISTIC REGRESSION MODEL. Suppose we have n observations, and the value of the dependent variable for observation i (y_i) is coded as either a 1 or 0. Assuming one's prior belief about the coefficient vector β is that it follows a normal distribution with mean b_0 and variance-covariance matrix B_0 , the posterior density of β is (up to a constant of proportionality) given by

$$p(\beta|y) \propto \prod_{i=1}^n \left(\frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)} \right)^{y_i} \times \left(1 - \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)} \right)^{1-y_i} \times \exp \left(-\frac{(\beta - b_0)' B_0^{-1} (\beta - b_0)}{2} \right).$$

To summarize this posterior density, we may want to calculate such quantities as the posterior mean of β and the posterior variance of β . Suppose we want to use the posterior mean of β (defined as $E[\beta|y] = \int \beta p(\beta|y) d\beta$) as a Bayesian point estimate. Doing the integration necessary to calculate the posterior mean is a nontrivial matter, even for such a simple model. Another approach to calculating the posterior mean of β is to use an MCMC algorithm to take a large random sample $\beta^{(1)}, \dots, \beta^{(m)}$ of size m approximately from $p(\beta|y)$ and to estimate the posterior mean of β as

$$\hat{E}[\beta|y] = \frac{1}{m} \sum_{i=1}^m \beta^{(i)}.$$

The following Metropolis algorithm in Figure 3 does just that.

```

initialize  $\beta^{(0)}$ 
for (i in 1 to m) {
  generate  $\beta_{can}^{(i)} \leftarrow \beta^{(i-1)} + z^{(i)}$ 
  calculate  $\alpha \leftarrow \frac{f(\beta_{can}^{(i)}|y)}{f(\beta^{(i-1)}|y)}$ 
  set  $\beta^{(i)} \leftarrow \begin{cases} \beta_{can}^{(i)} & \text{with probability } \min\{1, \alpha\} \\ \beta^{(i-1)} & \text{with probability } 1 - \min\{1, \alpha\} \end{cases}$ 
  store  $\beta^{(i)}$ 
}
calculate  $\bar{\beta} \leftarrow \frac{1}{m} \sum_{i=1}^m \beta^{(i)}$ 

```

Figure 3

where $z^{(i)}$ is a realization of a zero mean multivariate normal random variable with a fixed variance-covariance matrix drawn at iteration i , and $f(\beta|y)$ is any function that is proportional to $p(\beta|y)$. Note that, in practice, one would discard the first several hundred to several thousand samples of β as burn-in iterations to eliminate sensitivity to the starting value of β .

—Kevin M. Quinn

REFERENCES

- Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains. *Biometrika*, 57, 97–109.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1092.
- Robert, C. P., & Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer.

MATCHING

Matching refers to creating sets of TREATMENT and CONTROL units based on their similarity on one or more covariates. In randomized experiments, matching reduces error variance, thus increasing STATISTICAL POWER. In nonrandomized experiments, matching

increases similarity of treatment and control groups. Matching in nonrandomized experiments does not equate groups as well as randomization, but it can reduce bias.

Variations of matching include the following. *Matched-pairs* designs match a single control unit to a single treatment unit, yielding equal numbers of units per group. *Matched-sets* designs match one or more treatment units with one or more control units. Groups can be matched on a single variable (*univariate* matching) or several variables (*MULTIVARIATE* matching). *Exact* matching pairs treatment and control units that have the same value on the COVARIATE(S); however, finding an exact match for all units may be difficult. Exact multivariate matching is made easier by collapsing the matching variables into a single variable called a *propensity score*—the predicted probability of treatment or control group membership from a LOGISTIC REGRESSION. *Distance* matching matches pairs or sets of units to minimize the overall difference between control and treatment groups. Variations of exact and distance matching include full matching, optimal matching, and greedy matching (Shadish, Cook, & Campbell, 2002).

Examples of matching include the following: (a) Bergner (1974), who rank ordered 20 couples on pretest scores on a couple communication measure and then randomly assigned one couple from each pair to treatment or control conditions; (b) Cicerelli and Associates (1969), who compared children enrolled in Head Start to eligible but not enrolled children of the same gender, ethnicity, and kindergarten education; and (c) Dehejia and Wahba (1999), who used propensity score matching of those receiving labor training with controls drawn from a national survey. Matching in nonrandomized experiments can cause significant problems if not done carefully (Shadish et al., 2002).

—William R. Shadish and M. H. Clark

REFERENCES

- Bergner, R. M. (1974). The development and evaluation of a training videotape for the resolution of marital conflict. *Dissertation Abstracts International*, 34, 3485B. (UMI No. 73-32510)
- Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford.
- Cicerelli, V. G., & Associates. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development* (2 vols.). Athens: Ohio University and Westinghouse Learning Corp.

- Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053–1062.
- Rosenbaum, P. R. (1998). Multivariate matching methods. *Encyclopedia of Statistical Sciences*, 2, 435–438.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for general causal inference*. Boston: Houghton Mifflin.

MATRIX

A *matrix* can be defined as a rectangular array of real numbers or elements arranged in *rows* and *columns*. The *dimension* or *order* of a matrix is determined by its numbers of rows and columns. For example, **A** is said to be an m by n (denoted as $m \times n$) matrix because it contains $m \times n$ real numbers or elements arranged in m rows and n columns, as in

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & a_{ij} & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

where a_{ij} ($i = 1, 2, \dots, m; j = 1, 2, \dots, n$) is an *element* of matrix **A**.

Note that the first subscript for an element of a matrix—that is, a_{ij} —is the index for row and the second subscript for column, counting from the upper left corner. For example, a_{ij} is the value of the element in the i th row and the j th column. If a matrix **B** is one in which $a_{ij} = a_{ji}$ for all i and j , then **B** is called a *symmetric matrix*. For example,

$$\mathbf{B} = \begin{bmatrix} 1 & 4 & 7 \\ 4 & 6 & 5 \\ 7 & 5 & 9 \end{bmatrix}.$$

An $m \times 1$ matrix is called a *column vector* because it only has one column. Likewise, a $1 \times n$ matrix is called a *row vector* because it only has one row. A 1×1 matrix is simply a number and usually called a *scalar*.

A matrix that has an equal number of rows and columns is called a *square matrix*. For example, the symmetric matrix **B** illustrated above is a 3×3 square matrix. The sum of the diagonal elements

of a square matrix is called the *trace* of the square matrix. For example, $\text{trace}(B) = 16$. Several particular types of matrices are frequently used in statistical and ECONOMETRIC analysis.

- A *diagonal matrix* is a square matrix such that each element that does not lie in the main (or principal) diagonal (i.e., from the upper left corner to the lower right corner) is equal to zero. That is, $a_{ij} = 0$ if $i \neq j$. For example, $\mathbf{C}$ is an $m \times m$ diagonal matrix,

$$\mathbf{C} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & a_{mm} \end{bmatrix}.$$

- An *identity matrix* or a unit matrix is a special diagonal matrix whose main diagonal elements are all equal to 1. It is usually denoted $\mathbf{I}_m$, where m is the dimension or order of this identity matrix. (Thus, $\text{trace}(\mathbf{I}_m) = m$.) For example, $\mathbf{I}_3$ is an identity matrix,

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

- A *triangular matrix* is a particular square matrix that has only zeros either above or below the main diagonal. If the zeros are *above* the diagonal, the matrix is *lower triangular* because nonzero elements are below the diagonal. That is,

$$a_{ij} = \begin{cases} a_{ij}, & \text{for } i \geq j \\ 0, & \text{for } i < j \end{cases}.$$

For example, $\mathbf{D}$ is a lower triangular matrix,

$$\mathbf{D} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}.$$

In contrast, when the zeros are below the diagonal, the matrix is *upper triangular*.

- A diagonal matrix whose main diagonal elements are all equal is called a *scalar matrix*. For example, the variance-covariance matrix of a LINEAR REGRESSION error variable that follows all the assumptions of

classical linear REGRESSION analysis is a scalar matrix given by

$$\mathbf{\Omega} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}.$$

Each of these matrixes has its own properties that are often applied in STATISTICS and econometrics. For example, matrix operations and finding the *transpose* and *inverse* of a matrix are useful tools in statistical and econometric analysis. More information on these topics may be found in Namboodiri (1984) and Chiang (1984) for an elementary discussion and in Golub and Van Loan (1996) for an advanced discussion.

—Cheng-Lung Wang

REFERENCES

- Chiang, A. C. (1984). *Fundamental methods of mathematical economics* (3rd ed.). New York: McGraw-Hill.
- Golub, G. H., & Van Loan, G. F. (1996). *Matrix computations* (3rd ed.). Baltimore: Johns Hopkins University Press.
- Namboodiri, K. (1984). *Matrix algebra: An introduction*. Beverly Hills, CA: Sage.

MATRIX ALGEBRA

Manipulation of matrices can be used to solve complex relationships between several variables over many observations. Although such manipulations are carried out by computer programs for relationships beyond the most simple ones, the three basic MATRIX operations are the same and can be easily demonstrated.

MATRIX ADDITION

The matrices $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$ with equal order $m \times n$ may be added together to form the matrix $\mathbf{A} + \mathbf{B}$ of order $m \times n$ such that $\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]$. For example, if

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

and

$$\mathbf{B} = \begin{bmatrix} e & f \\ g & h \end{bmatrix},$$

then

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a + e & b + f \\ c + g & d + h \end{bmatrix}.$$

The commutative and associative laws both hold for matrix addition.

MULTIPLICATION BY A NUMBER

If $\mathbf{A} = [a_{ij}]$ is a matrix and k is a number, then the matrix $k\mathbf{A} = [ka_{ij}] = \mathbf{A}k$. For example, if

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

then

$$k\mathbf{A} = \begin{bmatrix} ka & kb \\ kc & kd \end{bmatrix}.$$

Note that every element of the matrix is multiplied by the number (unlike multiplying a determinant by a number). Multiplication by a number is commutative.

MATRIX MULTIPLICATION

Two matrices can be multiplied only if they are conformable. To be conformable, the number of columns in the first matrix must equal the number of rows in the second. For the matrices $\mathbf{A} = [a_{ik}]$ with order $m \times p$ and $\mathbf{B} = [b_{kj}]$ with order $p \times n$, the matrix $\mathbf{A} \times \mathbf{B}$ of order $m \times n$ is

$$\mathbf{AB} = \left[\sum_{k=1}^p a_{ik}b_{kj} \right].$$

Because the matrices must be conformable, the order of multiplication makes a difference (i.e., matrix multiplication is not commutative). For the resulting matrix $\mathbf{AB}$, we say that $\mathbf{B}$ is premultiplied by $\mathbf{A}$ or that $\mathbf{A}$ is postmultiplied by $\mathbf{B}$. For the case when $m = p = n = 2$, we would have

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

and

$$\mathbf{AB} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}.$$

Notice that each element of $\mathbf{AB}$ is the sum of the product of a particular element of $\mathbf{A}$ and a particular element of $\mathbf{B}$. Notice also that the inside number of each product pair is equal (because it is the k value) and that the numbers go from 1 to p (which equals 2 in this case) for each sum. For matrices of unequal size, the product matrices, $\mathbf{AB}$ and $\mathbf{BA}$, may be of a different order than either of the original matrices and may be of different order from each other. It is also possible that one or the other may simply not exist. For example, if $\mathbf{A}_{2 \times 4}$ and $\mathbf{B}_{4 \times 1}$, then $\mathbf{AB}_{2 \times 1}$, but $\mathbf{BA}$ would not exist because the columns in $\mathbf{B}$ do not equal the rows in $\mathbf{A}$. Note that even if $\mathbf{AB}$ and $\mathbf{BA}$ both exist and are of the same size, the two are not necessarily equal.

Following from these three operations, matrix subtraction is accomplished by multiplying the matrix to be subtracted by -1 and then adding the two matrices. Similarly, a form of matrix division requires finding the inverse of the intended divisor matrix and then multiplying that inverse with the second matrix. Calculating the inverse of a matrix involves a series of manipulations, and not all matrices have inverses. In particular, a matrix must be both square and nonsingular (i.e., a nonzero determinant) for it to have an inverse. Thus, if $\mathbf{A}$ is a square and nonsingular matrix, then its inverse, $\mathbf{A}^{-1}$, can be written as

$$\mathbf{A}^{-1} = \left(\frac{1}{|\mathbf{A}|} \right) \text{adj } \mathbf{A},$$

where $|\mathbf{A}|$ is the determinant of $\mathbf{A}$, and $\text{adj } \mathbf{A}$ is the adjoint of $\mathbf{A}$.

An important use of matrix algebra is in the calculation of REGRESSION coefficients. For example, the simple linear regression model may be expressed in matrix terms as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{Y}$ is the column vector of dependent variable values, $\mathbf{X}$ is the matrix of independent variable values, and $\boldsymbol{\beta}$ is the matrix of regression coefficients. Manipulation of the equation using matrix algebra results in the solution, $\boldsymbol{\beta} = \mathbf{X}^{-1}\mathbf{Y}$.

—Timothy M. Hagle

REFERENCES

- Cullen, C. G. (1990). *Matrices and linear transformations* (2nd ed.). New York: Dover.
- Pettoufrezzo, A. J. (1978). *Matrices and transformations*. New York: Dover.

MATURATION EFFECT

The *maturation effect* is any biological or psychological process within an individual that systematically varies with the passage of time, independent of specific external events. Examples of the maturation effect include growing older, stronger, wiser, and more experienced. These examples involve long-lasting changes. The maturation effect also can be transitory as in the case of fatigue, boredom, warming up, and growing hungry.

Campbell and Stanley (1963), in their classic *Experimental and Quasi-Experimental Designs for Research*, identified the maturation effect as one of eight rival hypotheses that could compromise the INTERNAL VALIDITY of an experiment. *Internal validity* is concerned with correctly concluding that an independent variable and not some extraneous variable is responsible for a change in the dependent variable. Internal validity is the sine qua non of experiments.

The simplest EXPERIMENTAL DESIGN that enables a researcher to rule out maturation and some other extraneous variables as rival explanations for a change in the dependent variable is the posttest-only control group design. Following Campbell and Stanley (1963), the design can be diagrammed as follows:

<i>R</i>	<i>X</i>	<i>O</i>
<i>R</i>		<i>O</i>

The *Rs* indicate that participants are randomly assigned to one of two conditions: a treatment condition, denoted by *X*, and a control condition, denoted by the absence of *X*. The control condition is administered to the participants as if it is the treatment of interest. The *O*s denote an observation or measurement of the dependent variable. The order of the events is *R* followed by *X* or the control condition and then *O*. The vertical alignment of the *Rs* indicates that the assignment of participants to the two groups occurs at the same time. Similarly, the vertical alignment of the *O*s indicates that the two groups are measured at the same time. RANDOM ASSIGNMENT is an essential feature of the design. It is used to achieve comparability of the participants in the two conditions prior to the administration of the independent variable. If the sample size is large and each participant has an equal chance of being assigned to a particular condition, a researcher can be reasonably confident that known as well as unsuspected extraneous variables are evenly distributed over the

conditions. Hence, maturation effects are controlled in the sense that they should be manifested equally in the two groups. The data for this simple design are often analyzed with a *t*-statistic for independent samples.

The posttest-only control group design can be modified for experiments that have *p* conditions, where *p* is greater than 2. The modified design in which $X_1, X_2, \dots, X_{p-1}$ denote treatment conditions and X_p denotes a control condition is as follows:

<i>R</i>	X_1	<i>O</i>
<i>R</i>	X_2	<i>O</i>
$\vdots$	$\vdots$	$\vdots$
<i>R</i>	X_{p-1}	<i>O</i>
<i>R</i>	X_p	<i>O</i>

The data for this experiment are usually analyzed using a completely randomized analysis of variance design.

—Roger E. Kirk

REFERENCE

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

MAXIMUM LIKELIHOOD ESTIMATION

At the close of the 19th century, Karl Pearson showed that by understanding statistical DISTRIBUTIONS, researchers had a powerful tool for better understanding outcomes from EXPERIMENTS and observational studies. Pearson went one step further, showing that by understanding a small number of CONSTANTS governing statistical distributions—that is, PARAMETERS—we could understand everything about the distributions. So long as we knew the parameters governing some distribution, we did not need to “see” the entire distribution. At the close of the 19th century, the field of parametric statistics was born.

This, of course, was a powerful insight. The problem remained, however, as to precisely how the empirical information from these experiments and observational studies could be harnessed to learn something about those distributional parameters. Pearson’s answer to that crucial question was based on collecting as many empirical observations as possible, hopefully from the entire POPULATION, and then

calculating characteristics of the empirical distribution—for example, the MEAN, VARIANCE, SKEWNESS, and KURTOSIS. These calculated empirical moments were in turn used to map the empirical distribution on to some known theoretical distribution, typically one of the so-called Pearson-Type distributions (Stuart & Ord, 1987).

At about the same time, a young genius by the name of Ronald Aylmer Fisher—and, by all accounts, Pearson's intellectual superior and to whom Pearson is well known to have been less than accommodating and at times openly hostile (Salsburg, 2001)—proved that Pearson's solutions to the ESTIMATION question were significantly flawed. Fisher showed that Pearson's estimates were oftentimes inconsistent and most always inefficient. By this, Fisher meant that Pearson's estimates for the parameters in a probability distribution were oftentimes not likely to converge in probability to the true population parameter, no matter how much DATA were collected (i.e., they were inconsistent), and that they have more variability from SAMPLE to sample than other ways of obtaining estimates (i.e., they were inefficient).

To replace Pearson's flawed methods with more sound estimation procedures, Fisher centered his attention on the probability of observing the data under different estimates of a distribution's parameters. More precisely, he sought to obtain, under general conditions, the parameter estimate that maximized the probability of observing the data that were collected or, as it is often called, the data likelihood. Not only were these *maximum likelihood estimates* superior in that they in fact maximized the data likelihood, but *maximum likelihood estimators* were also shown to be consistent and efficient and, at a later point, asymptotically unbiased under very general conditions and for a wide range of models and distributions. At the beginning of the 20th century, the method of maximum likelihood was born.

To better understand this method, it is useful to consider some RANDOM VARIABLE on observation i . Denote this random variable as Y_i and the data observed by the researcher as y_i . Suppose we wish to estimate the parameters governing the distribution from which Y_i originates. To do so, we need to know or, as is often the case, assume the form of the distribution of Y_i . To keep things general for the moment, denote the known or assumed distribution of Y_i as $f(Y_i; \theta)$, where θ is either the sole parameter or the vector of parameters governing the distribution of Y_i . In the case where Y_i is

discrete, $f(Y_i; \theta)$ gives the probability that Y_i will take on a specific value y_i and is often called the *probability function*. In the case where Y_i is continuous, $f(Y_i; \theta)$ governs the probability of observing a range of possible values y_i and is often called the PROBABILITY DENSITY FUNCTION.

Let's now assume that N data points $y_1, \dots, y_N$ are collected in some manner. Precisely how these data are collected (observational, experimental, etc.) is not so important so long as $y_1, \dots, y_N$ are *independent and identically distributed*, or *iid*. That is, we need to assume, for the data-generating mechanism giving rise to data point y_i and some other data point—say, $y_{i'}$ —that for all i and i' , (a) observing y_i is independent of observing $y_{i'}$, and (b) the probability (density) functions $f(Y_i; \theta)$ and $f(Y_{i'}; \theta)$ are the same. Of course, there are more or less complex ways to accommodate deviations from these ASSUMPTIONS in maximum likelihood estimation. But for now, let's ignore these additional considerations.

Given the iid assumption, the joint probability (density) function for the data—known as the *data likelihood* or, as Fisher called it, the probability of the sample—is given by the product of the individual probability (density) functions:

$$L(y; \theta) = \prod_{i=1}^N f(y_i; \theta). \quad (1)$$

The core of the idea that Fisher hit upon was to find values for θ that maximize equation (1), the likelihood of the data. Because products of runs of numbers present computing difficulties (underflows for small numbers, overflows for large numbers, general loss of precision, etc.), it is common to maximize instead the natural log of the data likelihood, or the log-likelihood,

$$\begin{aligned} \log\{L(y; \theta)\} &= \log\left\{\prod_{i=1}^N f(y_i; \theta)\right\} \\ &= \sum_{i=1}^N \log\{f(y_i; \theta)\}, \end{aligned} \quad (2)$$

as this gives the same solution as maximizing the FUNCTION in equation (1).

Aside from having the data, we must specify the form of $f(y_i; \theta)$ to find the value of θ that maximizes equations (1) and (2). This value is typically called the

maximum likelihood estimate and is most often denoted as $\hat{\theta}$. There are many distributions that may be used. Useful continuous distributions include the NORMAL, log-normal, LOGISTIC, beta, exponential, and gamma distributions; useful discrete distributions include the BINOMIAL, MULTINOMIAL, NEGATIVE BINOMIAL, and POISSON. Most introductory probability and statistics textbooks give a sampling of important distributions researchers may wish to consider (see, e.g., Hogg & Tanis, 1988). Eliason (1993) also provides examples in this regard. An important consideration when choosing an appropriate distribution is to consider the form of the observed y_i s in relation to the range allowed by a candidate distribution. For example, it may make little sense to choose a distribution where y_i is allowed to traverse the entire range of real numbers (as, for example, with the normal distribution) when the range of the observed data y_i is strictly positive (as, for example, with market wages) or is bounded between 0 and 1 (as, for example, with probabilities).

The normal distribution is by far the most commonly used distribution in the social sciences, either explicitly or, perhaps more often the case, implicitly. Moreover, it is common in didactic pieces such as these to assume a normal distribution and to then show that the maximum likelihood estimator for the mean of the distribution, often denoted μ , is the sample mean

$$\frac{1}{N} \sum_{i=1}^N y_i.$$

Although important, this is a well-known result, can be found in many sources, and, worst of all, can give an *incorrect* impression that these techniques must assume normality.

So, to make this more interesting and to show that all that is required is that we assume *some* distribution, let Y_i be distributed as a gamma random variable. The gamma distribution can take on many forms, although it is often considered in its chi-square-like form, that is, right-skewed with the distribution anchored on the left at zero. It is also the parent distribution to the exponential and the CHI-SQUARE. Perhaps more important from a social science perspective, the gamma distribution is interesting in that many reward and related distributions (e.g., wealth, wages, power, etc.) tend to follow a typical gamma distribution.

As considered in Eliason (1993), the two-parameter gamma distribution, with parameters $\mu = E\{Y_i\}$ (i.e., μ gives the expected mean value of Y_i) and

$\nu = E\{Y_i\}^2 / V\{Y_i\}$ (i.e., ν gives the ratio of the squared expected value over the variance of Y_i), is given by

$$f(y_i; \theta = [\mu, \nu]) = \left(\frac{1}{\Gamma\{\nu\}} \right) \left(\frac{\nu}{\mu} \right)^{\nu} y_i^{(\nu-1)} \exp\left\{ \frac{-\nu y_i}{\mu} \right\}. \quad (3)$$

To find the maximum likelihood solution for the parameter that governs the central tendency of this distribution, μ , let's first take the natural log of the gamma density in equation (3),

$$\log\{f(y_i; \theta = [\mu, \nu])\} = -\log \Gamma\{\nu\} + \nu \log\{\nu/\mu\} + (\nu - 1) \log\{y_i\} - (\nu y_i/\mu). \quad (4)$$

The log-likelihood for the data is then given by the sum of these natural logs,

$$\begin{aligned} \log\{L(y_i; \theta = [\mu, \nu])\} &= -N \log \Gamma\{\nu\} + \nu N \log\{\nu/\mu\} + (\nu - 1) \\ &\quad \sum_{i=1}^N \log\{y_i\} - (\nu/\mu) \sum_{i=1}^N y_i. \end{aligned} \quad (5)$$

The next step involves a working knowledge of first derivatives, as we need to calculate the first derivative of equation (5) with respect to the parameter of interest, μ , to solve for the maximum likelihood solution for μ . Although calculating first derivatives may be difficult, understanding first derivatives need not be. That is, a first derivative simply gives the rate of change of some function relative to something else. In this case, that function is the log-likelihood in equation (5), and the something else is the parameter μ , which is given by

$$\begin{aligned} &\frac{\partial \log\{L(y_i; \theta = [\mu, \nu])\}}{\partial \mu} \\ &= \left(\frac{\nu}{\mu^2} \right) \sum_{i=1}^N y_i - \left(\frac{\nu}{\mu} \right) N. \end{aligned} \quad (6)$$

From calculus, we know that when the first derivative (with respect to μ) equals zero, the corresponding value for μ is either at its maximum or minimum. In this case, we do have the maximum—to determine this requires a working knowledge of second

derivatives and is thus beyond the scope of this entry. Therefore, the last step involves setting the derivative in equation (6) to zero and solving for μ :

$$\begin{aligned} \left(\frac{\nu}{\mu^2}\right) \sum_{i=1}^N y_i - \left(\frac{\nu}{\mu}\right) N \\ = 0 \Rightarrow \hat{\mu} = \sum_{i=1}^N y_i / N. \end{aligned} \quad (7)$$

That is, the maximum likelihood estimator for μ , $\hat{\mu}$, for the gamma log-likelihood in equation (5) is the sample mean.

Importantly, these techniques are very general. Instead of μ , we could have been interested in the maximum likelihood solution of ν in the gamma, a component of the variance of the random variable Y_i . In addition, in the social sciences, we are often interested not only in the distributional parameters for some variable (say, e.g., the parameters governing the distributions of wages, prestige, or power) but in the way some other variables (say, e.g., attributes of individuals, jobs, markets, and institutions) come together to affect those parameters. This requires our attention to shift to an additional layer of the problem—to model specification and functional forms linking variables to parameters of interest.

To take an overly simplistic example, we may posit that attributes X_i (e.g., gender, race, age, degree attained, social class, etc.) of individuals embedded in labor market institutions with attributes Z_j (e.g., degree of labor market regulation, closed vs. open systems of jobs, degree of market concentration, unionization rate, etc.) interact to affect the distribution of wages. Note that here i indexes individuals, whereas j indexes markets. Recall that Pearson's breakthrough in the late 1800s showed us that the distribution of any variable is entirely governed by the parameters in that distribution. So, if we link our factors in an informative way (as expected by our theory of how individual and market attributes combine to affect wages) to the distributional parameters, we can show how these factors come together to affect the distribution as a whole.

Staying with the gamma distribution, we would thus link these factors to the parameters μ , which governs the central tendency of the wage distribution, and ν , which governs the shape of the wage distribution. Given that both μ and ν must be positive in the gamma distribution, the function linking these distributional parameters to the factors in X_i and Z_j must reflect

this. So, the simplest model adhering to that condition would result in a log-linear link with additive effects. That is,

$$\mu_{ij} = \exp\{\beta_0 + \beta_1 X_i + \beta_2 Z_j\}, \quad (8)$$

and

$$\nu_{ij} = \exp\{\gamma_0 + \gamma_1 X_i + \gamma_2 Z_j\}, \quad (9)$$

where β_1 , β_2 , γ_1 , and γ_2 are all vectors of model parameters with dimensions to accommodate the dimensions in X_i and Z_j . Maximum likelihood estimates for the β s and γ s would then contain the information as to how the individual and market factors in X_i and Z_j affect the wage distribution, given the model specification in equations (8) and (9).

To be consistent with the simple example of embeddedness given above, however, we would at least require an interaction that in part governs the distribution of wages. Thus, we would still have a log-linear link function, but now with an interaction term included:

$$\mu_{ij} = \exp\{\beta_0 + \beta_1 X_i + \beta_2 Z_j + \beta_3 X_i Z_j\}, \quad (10)$$

$$\nu_{ij} = \exp\{\gamma_0 + \gamma_1 X_i + \gamma_2 Z_j + \gamma_3 X_i Z_j\}. \quad (11)$$

In linking these factors in this way to the distributional parameters, the researcher can answer important questions regarding how individual and market attributes come together to affect the distribution of wages.

Although the models in equations (8) through (11) provide an additional layer of complexity, the basic ideas are not much different from those in the previous example in which we had no model. That is, given the data, we would (a) choose an appropriate probability (density) function for the outcome variable of interest, (b) create the functions linking the independent variables of interest to the parameters governing the probability (density) function for the outcome variable, and (c) maximize that log-likelihood to obtain the maximum likelihood estimates for the model parameters. Any statistical software, such as SAS's NLP procedure, that allows the user to write and maximize an arbitrary function may be used to obtain maximum likelihood estimates in this fashion. Once obtained, for the above example, these maximum likelihood estimates would provide the foundation for interpreting the effects of the individual and market attributes, as well as their interactions, on the distribution of wages.

The above represents the core ideas underlying the method of maximum likelihood as developed initially by R. A. Fisher. In current-day computer packages, many distributional and model parameters are estimated with maximum likelihood routines and algorithms, oftentimes behind the scenes and not obvious to the user of point-and-click statistical packages such as SPSS. Maximum likelihood is in fact the workhorse behind the estimation of many statistical models used in social science research, including STRUCTURAL EQUATION MODELS, LOGISTIC REGRESSION models, PROBIT regression models, TOBIT models, censored regression models, EVENT HISTORY models, LOG-LINEAR MODELS, GENERALIZED LINEAR MODELS, and so on. Thus, understanding the logic behind maximum likelihood methods gives one a powerful tool for understanding these many disparate models used by social scientists.

Finally, there are many useful sources to learn more about maximum likelihood estimation. Eliason (1993) provides an accessible introductory treatment to the logic and practice of this method. King (1989) also provides a wonderfully useful text. Gill's (2001) monograph shows how maximum likelihood undergirds much of the estimation of generalized linear models. In addition, Cramer's (1986) text provides useful insights into this method. For a more in-depth discussion, Stuart, Ord, and Arnold (1999) provide a highly useful volume detailing maximum likelihood methods in comparison to other estimation methods, such as Bayesian estimation techniques.

—Scott R. Eliason

REFERENCES

- Cramer, J. S. (1986). *Econometric applications of maximum likelihood methods*. Cambridge, UK: Cambridge University Press.
- Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. Newbury Park, CA: Sage.
- Gill, J. (2001). *Generalized linear models: A unified approach*. Thousand Oaks, CA: Sage.
- Hogg, R. V., & Tanis, E. A. (1988). *Probability and statistical inference*. New York: Macmillan.
- King, G. (1989). *Unifying political methodology: The likelihood theory of statistical inference*. New York: Cambridge University Press.
- Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: Holt.

Stuart, A., & Ord, K. (1987). *Kendall's advanced theory of statistics: Vol. 1. Distributional theory* (5th ed.). New York: Oxford University Press.

Stuart, A., Ord, K., & Arnold, S. (1999). *Kendall's advanced theory of statistics: Vol. 2A. Classical inference & the linear model* (6th ed.). New York: Oxford University Press.

MCA. See MULTIPLE CLASSIFICATION ANALYSIS

McNEMAR CHANGE TEST. See McNEMAR'S CHI-SQUARE TEST

McNEMAR'S CHI-SQUARE TEST

Also known as the McNemar change test, this is the nonparametric analogue to the *T*-TEST for related samples, for data that do not meet the requirements of that PARAMETRIC test. In particular, this test covers situations in which the dependent variable is NOMINAL data (i.e., frequencies for categories). This test is literally looking for a significant change in the category of subjects.

To illustrate, let us consider the psychology and sociology departments at Freedonia University. They decided to recruit students into a common first-year social sciences program, letting the students select their degree subject at the end of the year. There was some suspicion that the sociology department would benefit more than the psychology department from this process in terms of students changing their minds. The sociology department suggested that it be left to a trial year and that a check be made at the end of the year based on the number of students changing, presenting the psychology department with a test of their "faith" in statistics. At the beginning of the year, the departments randomly selected two groups of 24 students claiming a strong commitment to pursuing one degree or the other from those in the class of 250. At the end of the year, they compared this to what they selected for a department for the next year. The results are shown in Figure 1. Was the psychology department always going to lose out because it ended up with 16 students

		Beginning Choice		
		Psyc	Soc	
Totals ⇒	Before	24	24	Totals
End of Year Choice	Psyc	7 A	9 C	16
	Soc	17 B	15 D	32
				After ↓

Figure 1 Before and After First-Year Course Choices for Degrees

and the sociology department with 32, or was this just a difference that was within the realm of random fluctuation?

This could be considered an example of a design that consists of pretest and posttest/observations on two groups using a dichotomous variable, in which each subject belongs to one of two alternative groups. The “treatment” here was seemingly the first-year course, although it could be said that it was the collective experience of learning what the two disciplines and teaching staff were really like.

The really important students to this test are the ones who changed their minds—the 9 in cell *C* who went from sociology to psychology and the 17 in cell *B* who went from psychology to sociology. Although one would expect some changes, if the first year exercised no differential effect on the students, one would expect as many going one way as the other. In other words, the expected change would have been

$$E = \frac{B + C}{2}.$$

Obviously, not every sample would produce exactly this number, so we resort to a statistical test to see what would be an acceptable variation from this. As this is a matter of categories, the chi-square test is appropriate. If we use *B* and *C* as the observed frequencies, these can be substituted into the following equation for the chi-square ratio:

$$\chi^2_{\text{ratio}} = \sum_{i=1}^m \frac{(O_i - E_i)^2}{E_i},$$

which would become

$$\chi^2_{\text{ratio}} = \frac{\left(B - \frac{B+C}{2}\right)^2}{\frac{B+C}{2}} + \frac{\left(C - \frac{B+C}{2}\right)^2}{\frac{B+C}{2}},$$

which reduces to

$$\chi^2_{\text{ratio}} = \frac{(B - C)^2}{B + C}, \quad df = 1 \text{ (McNemar change test)}.$$

Thus, for the example above,

$$\chi^2_{\text{ratio}} = \frac{(9 - 17)^2}{9 + 17} = \frac{64}{26} = 2.46.$$

Because $\chi^2(1, .05) = 3.84$, the result was considered nonsignificant; in other words, this was a chance occurrence. It was argued that this scheme would ultimately not influence the balance of the number of students changing their minds over a year.

—Thomas R. Black

REFERENCES

- Black, T. R. (1999). *Doing quantitative research in the social sciences: An integrated approach to research design, measurement, and statistics*. London: Sage.
- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.
- Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

MEAN

The mean is one of the most commonly used ways of summarizing the values of a variable. The arithmetic mean is the sum of all the values of the variable divided by the number of observations—that is, $\sum x_i / n$, where x_i represents the *i*th value of the variable of interest, x , and n is the sample size. For example, Table 1 shows the incomes of a sample of 12 people. The sum of the incomes of the 12 people is \$3,970 and, when divided by the n of 12, gives a mean of \$331. The mean is commonly what is meant when people refer to the AVERAGE. The mean of x is often written as $\bar{x}$ and is used to represent the best estimate of x_i , to which improved estimates can be compared in, for example, ordinary least squares. The extent to which the mean effectively summarizes a distribution will depend on the levels of dispersion in the distribution and on the existence of a skewed distribution or of “outliers.” For example, in the notional data in Table 1, Person F has an income (\$944) that far exceeds the next highest income. If this outlying case is excluded, the mean falls to \$275, which might be considered to be a value that is more typical of the incomes as a whole.

Table 1 Net Incomes of a Sample of 12 People

<i>Person</i>	<i>Net Income (\$ per week)</i>
A	260
B	241
C	300
D	327
E	289
F	944
G	350
H	195
I	160
J	379
K	214
L	311
<i>Total</i>	<i>3,970</i>
<i>Mean</i>	<i>331</i>

Extreme values, over time, tend to move closer to the mean, as Galton first ascertained in his famous study of the relationship between fathers' and sons' heights (Galton, 1889). This is known as regression toward the mean and is the basis of regression analysis. The mean of a sample is, when combined with its CONFIDENCE INTERVALS, used to estimate the population mean for the given variable.

Less commonly used in the social sciences than the arithmetic mean are the geometric mean and the harmonic mean. The geometric mean is the n th root of the product of the data values, where n represents the number of cases. That is, it is given by

$$(x_1^* x_2^* \dots x_n)^{1/n}.$$

Another way of thinking about this is that the log of the geometric mean is the arithmetic mean of the logs of the data values, or

$$\log G = 1/n \left(\sum \log x \right).$$

In the example in Table 1, therefore, $\log G$ is equal to 2.47, and the geometric mean is \$296.5. The properties of the geometric mean make it more appropriate for measuring the mean in contexts in which the increase in values is itself geometrically related to the starting point. For example, geometric means are applied to price (or other sorts of) indexes, in which the prices are expressed as a ratio of the beginning of the series. Another example of an application can be found in estimating the size of a population between measurements at different time points by taking its mean. For example, the geometric mean of the populations from two consecutive decennial censuses would

provide an estimate of the population at the 5-year midpoint.

The harmonic mean, on the other hand, is used to summarize frequency data or data when results are grouped by outcome. The reciprocal of the harmonic mean is the arithmetic mean of the reciprocals of the data values. It is therefore given by

$$1/H = 1/n \sum 1/x \quad \text{or} \quad N \times 1/ \sum (1/x).$$

For example, if, in a study of 20 families, 7 had 1 child, 8 had 2 children, and 5 had 3 children, the harmonic mean of family size would be $1/(7 + 4 + 1.7)/20 = 1.6$ children. This is lower than the arithmetic mean, which would come to 1.9 children. In fact, the harmonic mean is always lower than the arithmetic mean. Because of the reciprocal nature of the calculation of the harmonic mean, it also represents the arithmetic mean when the relationship between the component frequencies is expressed in the reverse way. Thus, in this case, there is 1 family per child for 7 children, half a family per child for 8 children, and a third of a family per child for 5 children, which comes to .635 families per child or, standardizing per family, to 1.6 children per family. Thus, it can be used when one wishes to calculate the reciprocal arithmetic mean, for example, with means of prices when calculated via rates per dollar or via dollars for set quantities. How much lower it is than the arithmetic mean depends on the amount of dispersion and the size of the mean.

—Lucinda Platt

See also AVERAGE, CONFIDENCE INTERVALS, MEASURES OF CENTRAL TENDENCY

REFERENCES

Galton, F. (1889). *Natural inheritance*. London: Macmillan.
Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics: Vol. 1. Distribution theory* (5th ed.). London: Griffin.
Yule, G. U., & Kendall, M. G. (1950). *An introduction to the theory of statistics* (14th ed.). London: Griffin.

MEAN SQUARE ERROR

In REGRESSION, the mean square error is the total sum of squares minus the regression sum of squares,

adjusted for the sample size (Norusis, 1990, p. B-75). The adjustment takes account of the DEGREES OF FREEDOM as well as sample size. The mean squared error is closely involved in the calculation of statistics measuring GOODNESS OF FIT, including the R^2 and the F statistic for regression.

The mean square error in regression relates to the mean squared variation for a single variable as follows. The mean squared variation is a unit-free measure of the dispersion of a continuous variable. The mean squared variation takes no account of skewness. The equation that defines the mean squared variation allows for sample size and for the deviations of each observation from the mean.

If we denote the mean by

$$\bar{X} = \frac{\sum_i X_i}{n}$$

and notice that $(n - 1)$ is the degrees of freedom for this statistic, then the mean squared variation can be written as follows (see SPSS, 2001):

$$\text{Mean squared variation} = \frac{\bar{X}}{n - 1} = \frac{\sum_i X_i}{n \cdot (n - 1)}.$$

By comparison, in multivariate regression, the mean squared variation is the total sum of squared variations (TSS) divided by n , the sample size. However, when an estimate of a regression line using p variables has been obtained, the mean squared variation can be broken up into two parts: the variation explained by the regression and the residual variation. The latter is commonly known as the mean squared error. Define terms as shown below.

$$\text{Total sum of squares} = \text{TSS} = \sum (Y_i - \bar{Y})^2.$$

$$\text{The mean squared variation in total is } \frac{\text{TSS}}{n}.$$

$$\text{Regression sum of squares} = \text{RSS} = \sum (\hat{Y} - \bar{Y})^2.$$

The residual error, denoted here as ESS, is the difference between TSS and RSS, as shown as follows:

$$\text{ESS} = \sum (Y_i - \hat{Y})^2.$$

Each sum of squares term is then divided by its degrees of freedom to get the corresponding mean squares term.

The mean square of regression, or mean sum of squares (MSS), is

$$\frac{\text{TSS}}{k}.$$

The mean squared error, MSE, is

$$\frac{\text{ESS}}{n - k - 1}.$$

The ratio of these two is the F statistic for regression (Hardy, 1993, p. 22).

$$F = \frac{\text{MSS}}{\text{MSE}} = \frac{\frac{\text{TSS}}{k}}{\frac{\text{ESS}}{n-k-1}} = \frac{\text{TSS}}{\text{ESS}} \cdot \frac{n - k - 1}{k}.$$

Thus, the mean square error is crucial for deciding on the significance of the R^2 in multiple regression. Mean squares generally, however, play a role in a range of statistical contexts.

In the ANALYSIS OF VARIANCE, or ANOVA, the total mean squares represent the variation in the dependent variable, whereas the mean squares explained by different combinations of the independent variables are compared relative to this starting point.

Thus, in general, mean squares give a unit-free measure of the dispersion of a dependent variable. By comparing the total squared deviation with the regression sum of squares and adjusting for sample size, we use mean squares to move toward summary indicators of goodness of fit. Inferences from these measures (i.e., from R^2 , t -tests, and the F test) provide standard measures of the significance of inferences to the population.

It is important in regression analysis to adjust the resulting R^2 value; the ADJUSTED R^2 goes further in correcting the test statistic for the sample size. Its formula allows the estimator to be adjusted for the sample size, even beyond the corrections for degrees of freedom that give the mean squares of regression and the mean squared error.

Thus, although mean square error plays an important role in regression, further adjustments may be needed before inferences are drawn.

—Wendy K. Olsen

REFERENCES

- Hardy, M. (1993). *Regression with dummy variables* (Quantitative Applications in the Social Sciences Series, No. 93). London: Sage.

Norusis, M. (1990). *SPSS/PC + Statistics 4.0*. Chicago: SPSS.
 SPSS. (2001). *Reliability* (Technical section). Retrieved from
www.spss.com/tech/stat/algorithms/11.0/reliability.pdf

MEAN SQUARES

In general, mean squares provide a statistic that estimates the average squared deviations across the cases in a group while adjusting for the degrees of freedom of the estimate. Mean squares rest on an assumption of RANDOM SAMPLING for inferences to be valid. The mean squares are used to calculate other test statistics.

In regression, the MEAN SQUARE ERROR is the total sum of squares minus the regression sum of squares, adjusted for n (cases) and k (independent variables).

The mean squared error, MSE, is

$$\frac{ESS}{n - k - 1},$$

with ESS the residual error.

MSE gives a unit-free estimate of the error relative to the mean, taking into account both sample size and the degrees of freedom.

Two examples illustrate two usages of mean squares: first in the ANALYSIS OF VARIANCE and second in the analysis of REGRESSION.

In the analysis of variance, a table such as Table 1 may be obtained.

The ratio of two mean square statistics follows an F distribution, whose significance (under the relevant set of assumptions) can be obtained from a table or from a statistical package.

Thus, for a set of 13 measures of citizens' confidence, with $n = 415$ cases, ANOVA gives the results shown in Table 1.

In this table, the F statistic is the ratio of 20.0803 and .3257 (SPSS, 1990, p. 868). When the ratio of mean squares explained to mean squares unexplained is high, one would reject the hypothesis that the between-measures variation is not substantial. In other words, citizens' confidence does not only vary between individuals but also varies considerably between measures. Formulas for the analysis of variance allow the calculation of degrees of freedom in such cases.

The second example illustrates results from regression estimation using ordinary least squares. In a study of logged wage rates with 31 independent variables in the equation, including years of education, years of

Table 1 ANOVA Mean Squares Results for Citizens' Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic
Between people	669	414	1.6177	
Within people	1,859	4,980	.3733	
Between measures	241	12	20.0803	61.66*
Residual	1,617	4,968	.3257	
Total	2,529	5,394	.4688	

SOURCE: SPSS (1990, p. 868).

*Significant at < 1% level.

full-time work experience, gender, and other variables, Table 2 results in the following:

Table 2 Mean Squares Table for Regression

	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic	Significance
Regression	602	31	19.431	129	.000
Residual	549	3,643	0.151		
Total	1,152	3,674			

The table indicates that the regression mean square is huge relative to the mean squared error; R^2 adjusted is 52%. In the table, the first column is divided by the second to obtain the mean square. Then the regression mean square is divided by the residual mean square to give an F statistic with 31 and 3,643 degrees of freedom. The significance of this figure is near zero.

If a simpler regression were run, the degrees of freedom change, but the F statistic remains strong (see Table 3); R^2 is 38%. With smaller data sets, one may find that the larger regression has a low F statistic due to the loss of degrees of freedom.

Table 3 Revision of F Statistic Using Mean Squares for a Simpler Regression

	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic	Significance
Regression	435	10	43	222	.000
Residual	717	3,664	0.196		
Total	1,152	3,674			

The predictor variables used here were age, years of unemployment, education in years, hours worked < 30, years of family care leave, number of years of

part-time employment, degree of segregation (males as percentage of total in that occupation), number of years of full-time employment, and full-time years worked (squared). See Walby and Olsen (2002) for details of these variables. In this example, F is higher (222) for the simpler regression.

—Wendy K. Olsen

REFERENCES

- SPSS. (1990). *SPSS reference guide*. Chicago: SPSS.
- Walby, S., & Olsen, W. (2002). *The impact of women's position in the labour market on pay and implications for UK productivity*. Cabinet Office Women and Equality Unit, November. Retrieved from www.womenandequalityunit.gov.uk

MEASURE. See CONCEPTUALIZATION, OPERATIONALIZATION, AND MEASUREMENT

MEASURE OF ASSOCIATION

A measure of association is a statistic that indicates how closely two variables appear to move up and down together or to have a pattern of values that observably differs from a random distribution of the observations on each variable. Measures of association fall into two main classes: parametric and nonparametric.

Two cautions precede the measurement of bivariate relationships. First, the sampling used before data collection begins has spatial and temporal limits beyond which it is dangerous to draw inferences (Olsen, 1993; Skinner, Holt, & Smith, 1989). Second, evidence of a strong association is not necessarily evidence of a causal relation between the two things indicated by the variables.

Three main types of association can be measured: CROSS-TABULATION, ordinal relationships, and CORRELATION. For cross-tabulations with categorical data, the available tests include the Phi, Kappa, and Cramér's V coefficients; the McNemar change test; and the Fisher and CHI-SQUARE TESTS. When one variable is ordinal and the other categorical, one may use the MEDIAN TEST, the Wilcoxon-Mann-Whitney test, or KOLMOGOROV-SMIRNOV TEST, among others. When both variables are continuous and normally distributed, PEARSON'S CORRELATION COEFFICIENT offers an indicator of how strongly one can infer an association from the sample to the population. If both variables are ordinal, Spearman's rank-order correlation coefficient or Kendall's TAU b (τ_b) may be used. Kendall's τ_b has the advantage of being comparable across several bivariate tests, even when the variables are measured on different types of ordinal scale.

OPERATIONALIZATION decisions affect the outcomes of tests of association. Decisions about technique may influence the choice of level of measurement. In addition, the nature of reality may affect how things can be measured (e.g., categorical, ordinal, interval, continuous, and continuous normally distributed levels of measurement).

It is crucial to choose the right test(s) from among the available nonparametric and parametric tests. Choosing a test is done in three steps: operationalize both variables, decide on a LEVEL OF MEASUREMENT for each variable, and choose the appropriate test (see Siegel & Castellan, 1988).

An example illustrates the ordinal measurement of attitudes that lie on a continuous underlying distribution. The example illustrates the use of the paired t statistic; a nonparametric statistic such as the Kolmogorov-Smirnov test for nonnormal distributions would be an acceptable alternative (see Table 1).

Among individuals of various social backgrounds, 154 were randomly sampled and asked how well they

Table 1 The Paired t -Test in an Attitude Survey About Occupations in India

<i>Job or Occupation</i>	<i>Mean of Attitudes About Young Men Doing Job</i>	<i>Mean of Attitudes About Young Women Doing Job</i>	<i>Paired t Statistic for the Paired Values</i>
Buying and managing livestock	4.2	3.1	-13.8*
Garment stitching with a sewing machine	2.4	2.8	3.7*

SOURCE: Field survey, 1999, Andhra Pradesh, India.

NOTE: The preferences were recorded during one-to-one interviews as 1 = *strongly disapprove*, 2 = *disapprove*, 3 = *neutral*, 4 = *approve*, and 5 = *strongly approve*.

*Significant at the 1% level.

would hypothetically have liked a young woman (or a young man) to do a certain type of job or training. Their answers were coded on a 5-point Likert scale that was observed to have different shapes for different jobs. The *t*-test measures the association between each pair of Likert scales: preference for a young man doing that job versus preference for a young woman doing that job. The paired *t*-test is often used in repeated-measures situations.

The local norms were that young men should do livestock management, whereas young women were more suited to garment stitching. As usual with bivariate tests, the direction of causality has not been ascertained, and the strong observed association could even be spurious.

—Wendy K. Olsen

REFERENCES

Olsen, W. K. (1993). Random samples and repeat surveys in South India. In S. Devereux & J. Hoddinott (Eds.) *Fieldwork in Developing Countries* (pp. 57-72). Boulder, CO: Lynne Rienner.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.). (1989). *Analysis of complex surveys*. Chichester, UK: Wiley.

MEASURES OF CENTRAL TENDENCY

Measures of central tendency are ways of summarizing a distribution of values with reference to its central point. They consist, primarily, of the MEAN and the MEDIAN. The MODE, although not strictly a measure of central tendency, also provides information about the distribution by giving the most common value in the distribution. It therefore tells us something about what is typical, an idea that we usually associate with a central point. Table 1 shows the number of dependent children in seven notional families. We can summarize the midpoint of this distribution with reference to the arithmetic mean (the sum of the values divided by the number of the observations), the median (or midpoint of the distribution), or the mode (the most common value).

Table 2 identifies the values of each of these three measures of central tendency as applied to the family

Table 1 Number of Children in Seven Families

Family	Number of Children Younger Than Age 16
Jones	1
Brown	1
Singh	1
Roberts	2
O'Neill	2
Smith	3
Phillips	6

Table 2 Measures of Central Tendency Derived From Table 1

Measure	Result From Table 1
Mean	16/7 = 2.3
Median	2
Mode	1

sizes given in Table 1. It shows that each measure produces a different figure for summarizing the center of the distribution. In this case, the limited range of the variable and the observations means that the differences are not substantial, but when the range of values is much wider and the distribution is not symmetrical, there may be substantial differences between the values provided by the different measures.

Although the mean can only be used with interval data, the median may be used for ordinal data, and the mode can be applied to nominal or ordinal categorical data. On the other hand, when data are collected in continuous or near-continuous form, the mode may provide little valuable information: Knowing that, in a sample of 100 people, two people earn \$173.56, whereas all the other sample members have unique earnings, tells us little about the distribution.

An additional, though rarely used, measure of central tendency is provided by the midrange, that is, the halfway point between the two ends of distribution. In the example in Table 1, the midrange would have the value 2.5. A modification of this is the midpoint between the 25th and 75th percentile of a distribution (i.e., the midpoint of the INTERQUARTILE RANGE).

Means can be related to any fixed point in the distribution, although they are usually assumed to be related to 0. For example, IQ tests set the standard for intelligence at 100. An average score for a class of students could be taken from the amount they differ from 100. The final average difference would then be added to (or subtracted from) 100. Such nonzero reference points are often taken as those in which the frequency of

observations is greatest, as this can simplify the amount of calculation involved.

HISTORICAL DEVELOPMENT

Measures of central tendency were in use long before the development of statistics. That is because, by telling us about the middle of the distribution, they tell us something about what is “normal” and how to place ourselves in relation to them. People construct ideas of how tall, poor, thin, or clever they are by reference to a perceived midpoint. Arithmetic means were long used as a check on the reliability of repeat observations; however, it was not until the 19th century that the possibilities of measures of central tendency for statistical science began to be developed. Francis Galton first explored the utility of the midpoint of the interquartile range as a summary measure, and he also developed Adolphe Quetelet’s work on the “average man” with investigations of deviation from both the mean and the median in his explorations of heredity (see AVERAGE). Francis Edgeworth explored the relevance of different means and the median to indexes, such as price indexes. Measures of central tendency remain critical to descriptive statistics and to understanding distributions, as well as to key statistical concepts such as regression.

APPLICATION OF MEASURES OF CENTRAL TENDENCY

Different measures of central tendency may be more relevant or informative in particular situations or with particular types of data. In addition, in

combination, they may give us added information about a distribution. In a symmetrical distribution, the mean and the median will have the same value. However, if the distribution is skewed to the right (this is usually the case in income distributions, for example), then the median will have a lower value than the mean. If the distribution is skewed to the left (as, for example, with exam results), then the value of the mean will be lower than that of the median.

Table 3 illustrates a small data table of the characteristics across five variables of 15 British households in 1999, derived from an annual national household survey. The measures that are appropriate to summarize location, or central tendency, vary with these variables. For example, when looking at the number of rooms, we might be interested in the mean, the median, or the mode. It is worth noting, in this case, that the mean provides us with a number (5.1) that cannot, because it is fractional, represent the actual number of rooms of any household. It is the same apparent paradox that arises when we talk of the average number of children per family being 1.7, which conjures up notions of subdivided children. For housing tenure (1 = *owned outright*, 2 = *buying with a mortgage*, 3 = *renting*), only the mode is appropriate as this is a categorical variable without any ordering. By contrast, although the local tax bracket is also a categorical variable, it is ordered by the value of the property, and thus the median may also be of interest here. For the income values, the mean or the median may be pertinent (or both). It is left to the reader to calculate the relevant measures of central tendency for each of the variables.

Table 3 The Individual Characteristics of 15 Households, Britain 1999

<i>Region</i>	<i>Number of Rooms</i>	<i>Housing Tenure</i>	<i>Household Income</i> (£ per week)	<i>Local Tax Bracket</i>
West Midlands	6	2	202.45	1
West Midlands	4	1	301	3
West Midlands	6	3	192.05	2
West Midlands	6	1	359.02	4
West Midlands	3	3	209.84	1
West Midlands	4	3	491.8	1
West Midlands	5	2	422.7	2
East Midlands	9	1	434.43	5
East Midlands	4	3	353	1
East Midlands	6	2	192	1
East Midlands	6	2	490.86	3
East Midlands	5	2	550	4
East Midlands	3	2	372.88	3
East Midlands	3	3	303	2
East Midlands	7	2	693.87	5

Finally, we see that households come from one of two regions—East Midlands or West Midlands. We might be interested to consider whether there is systematic difference between the midpoints of the variables according to region. Applications often require the comparison for different groups within a sample or for different samples; the *t*-TEST provides a way of comparing means and ascertaining if there are significant differences between them, whereas the MEDIAN TEST is used to compare medians across samples.

—Lucinda Platt

See also AVERAGE, INTERQUARTILE RANGE, MEAN, MEDIAN, MEDIAN TEST, MODE, *T*-TEST

REFERENCES

Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics: Vol. 1. Distribution theory* (5th ed.). London: Griffin.
Yule, G. U., & Kendall, M. G. (1950). *An introduction to the theory of statistics* (14th ed.). London: Griffin.

MEDIAN

The median is the midpoint of the distribution of a variable such that half the distribution falls above the median and half falls below it. That is,

$$\text{Median} = (n + 1)/2.$$

If the distribution has an even number of cases, then the median is deemed to fall precisely between the two middle cases and is calculated as the arithmetic MEAN of those two cases. For example, in the set of cases

1, 2, 5, 6, 8, 11, 14, 17, 45,

the median value is 8. In the set of cases

1, 5, 6, 12, 18, 19,

however, the median is $(6 + 12)/2 = 9$.

When the data do not have discrete values but are grouped either in ranges or with a continuous variable (e.g., age is given whereby each year actually represents the continuous possible ages across that period), a different approach is used to calculate the median. For example, Table 1 shows the frequencies of children in a small specialist school who fall into particular age bands.

Table 1 Distribution of Children by Age Ranges in a Small Specialist School (constructed data)

<i>Age Range (Years)</i>	<i>Frequency</i>	<i>Cumulative Frequency</i>
3 to 5.9	16	16
6 to 8.9	35	51
9 to 11.9	62	113
12 to 14.9	63	176
15 to 17.9	21	197

There are 197 children, so the median value will be the age of the 99th child. The 99th child falls in the 9 to 11.9-year age bracket. In fact, it falls 48 observations into that bracket and so can be counted as the 48/62 share of that bracket of 3 years added on to the age of 8.9 that it has already passed. That is, the median can be calculated as

$$8.9 + ((48/62) \cdot 3) = 11.2.$$

The median is, therefore, one of several MEASURES OF CENTRAL TENDENCY, or location, that can be used to understand aspects of the distribution. It is also a form of the AVERAGE, even if that term is normally reserved in common parlance for the arithmetic mean. Although the median does not have as many applications as the mean, it has certain advantages in giving an understanding of a distribution. The principal advantage is that it is not as susceptible to OUTLIERS—that is, extreme, potentially erroneous and potentially distorting cases—that will influence the mean value but not the median (see MEAN). For example, in the first example above, the value of 45 would appear to be such an outlier. Similarly, the median may form a valuable average when the final value of a distribution is unknown, rendering the mean impossible to calculate. The median can also be used for ordinal data, that is, data for which there is a hierarchical order to categories but the distances between the categories are not equal. An example is measures of educational achievement, ranked 1 for no qualifications, 2 for midlevel qualifications, 3 for higher qualifications, and so on. In such cases, it is meaningless to calculate the mean as the values do not represent numeric quantities. However, the median may be of interest in such instances.

The median can also be considered the 50th percentile and, as such, is a special instance of percentiles as measures of distribution. Percentiles are the values at which the named proportion of cases falls below the value and the remainder fall above it. Thus, the

25th percentile (or lower quartile) represents the value at which 25% of cases are of smaller value and the remaining 75% are larger. The median is thus part of a general system of describing a distribution by the values that break it into specified divisions. There are 99 percentiles in a distribution that divide it into 100 parts, 9 deciles (of which the median is the 5th) that divide it into 10 parts, and 3 quartiles (of which the median is the central one) that divide it into 4 parts.

—Lucinda Platt

See also AVERAGE, MEASURES OF CENTRAL TENDENCY

REFERENCES

- Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics: Vol. 1. Distribution theory* (5th ed.). London: Griffin.
- Yule, G. U., & Kendall, M. G. (1950). *An introduction to the theory of statistics* (14th ed.). London: Griffin.

MEDIAN TEST

The median test is a NONPARAMETRIC test to ascertain if different samples have the same location parameter, that is, the same *median*. That is, do the two samples come from the same population, with any differences between their medians being attributable to sampling variation? It can be generalized to other percentiles such as the 25th percentile (the first quartile) or the 75th percentile (the third quartile). Its origins have been dated to Alexander Mood's work on nonparametric methods from 1950 (see Mood & Graybill, 1963). It belongs to a group of simple linear rank statistics as they are concerned with the rank of an ordered distribution (see also WILCOXON TEST and KRUSKAL-WALLIS *H* TEST).

In the median test, a score is derived from the number of cases in the different samples that fall above the median (or other percentile) of the pooled sample. These are then compared with the score that would be expected under the null hypothesis (i.e., if the two samples belonged to the same population). The resulting test statistic has a CHI-SQUARE DISTRIBUTION, and its significance relative to the degrees of freedom can thus be calculated.

Table 1 Scores for Men and Women on Duration of Unemployment Relative to the Median

		Male	Female	Total
Duration of unemployment	> Median	55	35	90
	≤ Median	44	70	114
Total		99	105	204

For example, let us consider an ordinal variable giving months' duration of unemployment in a British sample of 204 working-age adults who experienced unemployment at some time before the mid-1980s. The question is, then, whether the median duration of unemployment is the same for both men and women. For the pooled samples, the median value is 2, indicating between 6 months and a year of unemployment. Table 1 shows the numbers of men and women whose duration falls above this median and those falling below or equal to it.

The test gives us a CHI-SQUARE TEST statistic of 10.2 for 1 degree of freedom. This is significant at the .001 level, which indicates that the case for independence is weak and that we can reject the null hypothesis. That is, it is very unlikely that the median duration of unemployment for men and women is the same.

—Lucinda Platt

REFERENCES

- Hajek, J., & Sidák, Z. (1967). *Theory of rank tests*. New York: Academic Press.
- Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York: John Wiley.
- Mood, A. M., & Graybill, F. A. (1963). *Introduction to the theory of statistics* (2nd ed.). New York: McGraw-Hill.

MEDIATING VARIABLE

A mediating variable is one that lies intermediate between independent causal factors and a final outcome. Like MODERATING VARIABLES, mediating variables are seen as intervening factors that can change the impact of *X* on *Y*. Mediating variables aim to estimate the way a variable *Z* affects the impact of *X* on *Y*, and they can be estimated in one of four ways: through PATH ANALYSIS, STRUCTURAL EQUATION MODELING, proxies embedded in linear regression, and MULTILEVEL MODELING.

In PATH ANALYSIS, separate regressions are run for different causal processes. Recursive regressions are those that invoke variables used in another regression (and instrumental variables regression is a case in point). The X variables in one regression have standardized beta coefficients, defined as

$$\beta_i = \mathbf{B}_i \cdot \frac{\mathbf{S}_{X_i}}{\mathbf{S}_Y},$$

which lead toward the final outcome Y . However, in addition, a separate equation for the mediating variable Z measures the association of elements of the vector $\mathbf{X}$ with Z , and Z in turn is a factor associated with the final outcome Y .

Two examples illustrate the possible impact of Z .

1. In the analysis of breast cancer, social behavior patterns such as smoking and the decision to start or stop taking the contraceptive pill influence the odds of cancer developing, whereas genetic and biological factors form another path to the onset of cancer (Krieger, 1994). These two sets of factors then interact and can offset each other's influence when we examine the odds of the patient dying from cancer. A complex path diagram results (see Figure 1).

The biologist would consider the social factors "mediating variables."

2. In the analysis of how attitudes are associated with final outcomes, attitudes may play a role mediating the direct effect of a sociodemographic characteristic on an outcome. Gender, for instance, was not directly a factor influencing eating-out behaviors but was a strong mediating factor. Women were more likely to wish they could eat out more often compared with men. Gender

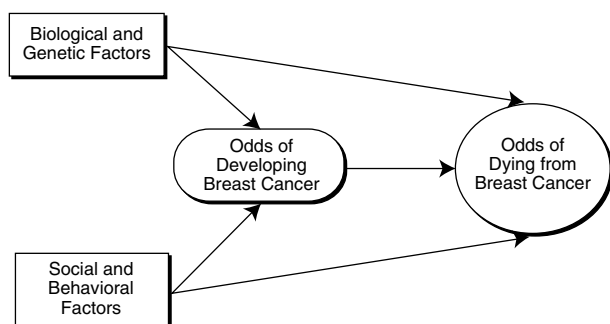


Figure 1 Causal Paths to Cancer Deaths

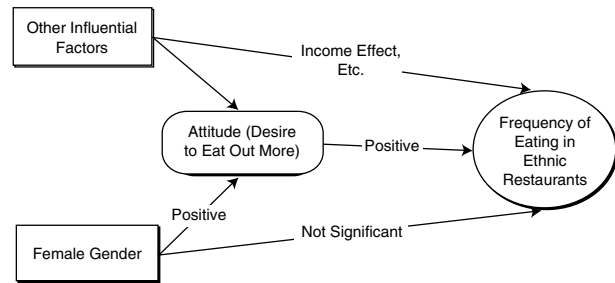


Figure 2 Causal Paths to Eating Out Frequently

thus mediated between the impact of income and other wealth indicators on the overall curiosity of a person regarding eating in ethnic restaurants (Olsen, 1999; Warde, Martens, & Olsen, 1999). See Figure 2.

Path analysis presents the error term as an unexplained causal process (Bryman & Cramer, 1997, chap. 10). In some economic analyses, the error term is presumed to represent a set of unrecorded causal factors, some of which are inherently unobservable, whereas in sociology, the error term might be thought to represent individual variation relative to a basic structuralist model.

Recent developments in the estimation of multilevel models using STATA software allow contextual effects to be represented through models in which the distribution to which a case is allocated is subject to conditional probabilities (allowing the contingent nature of contextual effects to be modeled) and is linked through a variety of functions (as in standard multilevel modeling) (Rabe-Hesketh, Pickles, & Skrondal, 2001). Thus, over time, new modeling structures for moderating factors become possible.

—Wendy K. Olsen

REFERENCES

- Arbuckle, J. (1997). *Amos users' guide version 3.6*. Chicago: Smallwaters Corporation.
- Bryman, A., & Cramer, D. (1997). *Quantitative data analysis with SPSS for Windows: A guide for social scientists*. New York: Routledge.
- Krieger, N. (1994). Epidemiology and the web of causation: Has anyone seen the spider? *Social Science & Medicine*, 39(7), 887–903.
- Olsen, W. K. (1999). *Path analysis for the study of farming and micro-enterprise* (Bradford Development Paper No. 3).

- Bradford, UK: University of Bradford, Development and Project Planning Centre.
- Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2001). *GLLAMM manual technical report 2001/01*. London: King's College, University of London, Department of Biostatistics and Computing, Institute of Psychiatry. Retrieved from www.iop.kcl.ac.uk/IoP/Departments/BioComp/programs/manual.pdf
- Warde, A., Martens, L., & Olsen, W. K. (1999). Consumption and the problem of variety: Cultural omnivorousness, social distinction and dining out. *Sociology*, 30(1), 105–128.

MEMBER VALIDATION AND CHECK

Also called *member check* and *respondent validation*, member validation is a procedure largely associated with QUALITATIVE RESEARCH, whereby a researcher submits materials relevant to an investigation for checking by the people who were the source of those materials. The crucial issue for such an exercise is how far the researcher's understanding of what was going on in a social setting corresponds with that of members of that setting. Probably the most common form of member validation occurs when the researcher submits an account of his or her findings (such as a short report or interview transcript) for checking.

The chief problem when conducting a member validation exercise is deciding what is to be validated. In particular, there is the issue of how far the submitted materials should include the kinds of interpretation that a researcher engages in when writing up findings for a professional audience. Even if members are able to corroborate materials with which they are submitted, so that they can confirm the researcher's interpretations of their interpretations, it is unreasonable to expect them to do so in relation to the social-scientific rendering of those interpretations. Furthermore, there is a difficulty for the researcher in knowing how best to handle suggestions by members that there has been a failure to understand them. In many cases, if this occurs, all that is needed is for the researcher to modify an interpretation. Sometimes, however, the lack of corroboration may be the result of members disliking the researcher's gloss. Indeed, the presentation of interpreted research materials to members may be the starting point for defensive reactions and wrangles over interpretations.

On the basis of their experiences of using member validation, Emerson and Pollner (1988) have pointed

to several other difficulties with its use: problems arising from a failure to read all of a report and/or from misunderstanding what was being said; difficulty establishing whether certain comments were indicative of assent or dissent; members, with whom the authors had been quite close, being unwilling to criticize or inadvertently asking member validation questions with which it would be difficult to disagree; and fears among members about the wider political implications of agreement or disagreement. Bloor (1997) has similarly argued that it is sometimes difficult to get members to give the submitted materials the attention the researcher would like, so that indifference can sound like corroboration.

Although member validation is clearly not an unproblematic procedure, as Lincoln and Guba (1985) observe, it can be crucial for establishing the credibility of one's findings and can also serve to alleviate researchers' anxieties about their capacity to comprehend the social worlds of others. It can be employed in connection with most forms of qualitative research. However, it is crucial to be sensitive to the limits of member validation when seeking such reassurance.

—Alan Bryman

See also RESPONDENT VALIDATION

REFERENCES

- Bloor, M. (1997). Addressing social problems through qualitative research. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (pp. 221–238). London: Sage.
- Emerson, R. M., & Pollner, M. (1988). On the use of members' responses to researchers' accounts. *Human Organization*, 47, 189–198.
- Lincoln, Y. S., & Guba, E. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

MEMBERSHIP ROLES

ETHNOGRAPHERS must assume social roles that fit into the worlds they study. Their access to information, their relationships to subjects, and their sociological interpretations are greatly influenced by these roles' character. Early Chicago school proponents advocated various research postures, such as the complete observer, observer-as-participant, participant-as-observer, and complete participant (Gold, 1958), ranging from least to most involved in the group. These

roles were popular when ethnographers were advised to be objective, detached, and distanced.

By the 1980s, ethnography partly relinquished its more OBJECTIVIST leanings, embracing its subjectivity. Ethnographers valued research roles that generated more depth in their settings. They sought positions offering regular, close contact with participants. Critical to these roles is their *insider* affiliation; when assuming *membership roles* in field research, individuals are not detached outsiders but instead take on participant status (Adler & Adler, 1987). This can vary between the *peripheral*, *active*, and *complete* membership roles.

Most marginal and least committed, the peripheral role still requires membership in the group's social world. Peripheral-member researchers seek an insider's perspective and direct, firsthand experience but do not assume functional roles or participate in core group activities. In our research on upper-level drug dealers and smugglers, we joined the social community and subculture, becoming accepted in the friendship network. We socialized with members, housed them, babysat their children, and formed long-term friendships. Yet, crucially, we refrained from participating in their drug trafficking (see Adler, 1985).

Less marginal and more central in the setting, active-member researchers take part in the group's core activities and assume functional and not merely social or research roles. Beyond sharing insider status, they interact with participants as colleagues. In our research on college athletics, Peter became an assistant coach on a basketball team, serving as the counselor and academic adviser. He regularly attended practices and functions, traveled with the team, and assisted in its routine functioning. However, his identity and job could detach from his coaching, and his professorial role remained primary (see Adler & Adler, 1991).

The complete membership role incorporates the greatest commitment. Researchers immerse themselves fully in their settings, are status equals with participants, and share common experiences and goals. Ethnographers may assume this role by selecting settings where they are already members or by converting. Although this position offers the deepest understanding and the least role pretense, it more frequently leads to conflicts between individuals' research and membership role demands. In gathering data on preadolescents' peer culture, we studied our children, their friends, and our neighbors, friends, schools, and community. Our parental membership role gave us naturally occurring

access to children's social networks and "after-school" activities. We used our membership relationships to recruit interviews and to penetrate beyond surface knowledge. At times, though, we were torn between wanting to actively parent these youngsters and needing to remain accepting to keep open research lines of communication (see Adler & Adler, 1998).

Membership roles offer ethnographers, to varying degrees, the deepest levels of understanding in their research.

—Patricia A. Adler and Peter Adler

REFERENCES

- Adler, P. A. (1985). *Wheeling and dealing*. New York: Columbia University Press.
- Adler, P. A., & Adler, P. (1987). *Membership roles in field research*. Newbury Park, CA: Sage.
- Adler, P. A., & Adler, P. (1991). *Backboards and blackboards*. New York: Columbia University Press.
- Adler, P. A., & Adler, P. (1998). *Peer power*. New Brunswick, NJ: Rutgers University Press.
- Gold, R. L. (1958). Roles in sociological field observations. *Social Forces*, 36, 217–223.

MEMOS, MEMOING

Memos are researchers' records of analysis. Although they take time to write, they are worth doing. Memos are a way of keeping the researcher aware and the research honest. All too often, researchers take shortcuts to the analytic process, writing notes in margins of transcripts, trusting their memories to fill in the details later when writing. Unfortunately, memories often fail, and the "golden nuggets" of insight that emerged during analysis are lost. Furthermore, memos are useful for trying out analytic ideas and working out the logic of the emergent findings. As memos grow in depth and breath, researchers can use them to note areas that need further investigation to round out the findings. When working in teams, memos can be used to keep everyone informed and to share ideas (Huberman & Miles, 1994). During the writing stage, researchers only have to turn to their memos to work out an outline.

There are no hard-and-fast rules for writing memos. However, here are a few suggestions for improving the quality of memos and enhancing their usefulness. Memos should be titled and dated. They should be analytic rather than descriptive, that is, focused on

concepts that emerged during the analysis and not filled with raw data (Dey, 1993). With the computer programs that are available, one can mark the raw data from which concepts emerged and refer back to the items when necessary (Richards & Richards, 1994). Memos should ask questions of the data, such as, "Where do I go from here to get more data on this concept, or what should I be thinking about when doing my next observation or interview?" As more data about a concept emerge, memos should explore concepts in terms of their properties and dimensions, such as who, what, when, where, how, why, and so on. The researcher should keep a piece of paper and a pencil available at all times, even by the bedside. One never knows when that sudden insight will appear, when the pieces of the puzzle will fall into place. Analysts should not let a lot of time elapse between doing the analysis and writing the memos. Not only do thoughts get lost, so does momentum. Analysts should not be afraid to be creative and write down those insights, even if they seem far-fetched at the time. One can always toss them out if they are not verified by data. Finally, memos can be used to keep an account of one's own response to the research process, including the personal biases and assumptions that inevitably shape the tone and outcome of analysis. This record keeping will prove useful later when writing.

Writing memos stimulates thinking and allows one to interact with data in ways that foster creativity while staying grounded at the same time. At the beginning of a research project, memos will appear more exploratory and directive than analytic. Often, they seem foolish to the researcher when looking back. This is to be expected. Doing research is progressive and an act of discovery, a process one can only keep track of in memos.

—Juliet M. Corbin

REFERENCES

- Dey, I. (1993). *Qualitative data analysis*. London: Routledge Kegan Paul.
- Huberman, A. M., & Miles, M. B. (1994). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 428–444). Thousand Oaks, CA: Sage.
- Richards, T. J., & Richards, L. (1994). Using computers in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 445–462). Thousand Oaks, CA: Sage.

META-ANALYSIS

The term *meta-analysis* refers to the statistical analysis of results from individual studies for purposes of integrating the findings. The person who coined the term in 1976, Gene Glass, felt that the accumulated findings of studies should be regarded as complex data points that were no more comprehensible without statistical analysis than the data points found in individual studies.

Procedures for performing meta-analysis had appeared in statistics texts and articles long before 1976, but instances of their application were rare. It took the expanding evidence in the social sciences and the growing need for research syntheses to provide impetus for the general use of meta-analysis. The explosion in social science research focused considerable attention on the lack of standardization in how scholars wishing to integrate literatures arrived at general conclusions from series of related studies. For many topic areas, a separate verbal description of each relevant study was no longer possible. One traditional strategy was to focus on one or two studies chosen from dozens or hundreds. This strategy failed to portray accurately the accumulated state of knowledge. First, this type of selective attention is open to confirmatory bias: A particular synthesist may highlight only studies that support his or her initial position. Second, selective attention to only a portion of all studies places little or imprecise weight on the volume of available tests. Finally, selectively attending to evidence cannot give a good estimate of the strength of a relationship. As evidence on a topic accumulates, researchers become more interested in "how much" rather than simply "yes or no."

Traditional synthesists also faced problems when they considered the variation between the results of different studies. Synthesists would find distributions of results for studies sharing a particular procedural characteristic but varying on many other characteristics. They found it difficult to conclude accurately whether a procedural variation affected study outcomes because the variability in results obtained by any single method meant that the distributions of results with different methods would overlap.

It seemed, then, that there were many situations in which synthesists had to turn to quantitative synthesizing techniques, or meta-analysis. The application of quantitative inference procedures to research synthesis

was a necessary response to the expanding literature. If statistics are applied appropriately, they should enhance the validity of synthesis conclusions. Meta-analysis is an extension of the same rules of inference required for rigorous data analysis in primary research.

HISTORICAL DEVELOPMENT

At the beginning of the 20th century, Karl Pearson (1904) was asked to review the evidence on a vaccine against typhoid. He gathered data from 11 relevant studies, and for each study, he calculated a CORRELATION coefficient. He averaged these measures of the treatment's effect across two groups of studies distinguished by their outcome variable. On the basis of the average correlations, Pearson concluded that other vaccines were more effective. This is the earliest known quantitative research synthesis.

Three quarters of a century after Pearson's research synthesis, Robert Rosenthal and Donald Rubin (1978) undertook a synthesis of research studying the effects of interpersonal expectations on behavior in laboratories, classrooms, and the workplace. They found 345 studies that pertained to their hypothesis. Nearly simultaneously, Gene Glass and Mary Lee Smith (1978) conducted a review of the relation between class size and academic achievement. They found 725 estimates of the relation based on data from nearly 900,000 students. This literature revealed 833 tests of the treatment. John Hunter, Frank Schmidt, and Ronda Hunter (1979) uncovered 866 comparisons of the differential validity of employment tests for Black and White workers.

Largely independently, the three research teams rediscovered and reinvented Pearson's solution to their problem. They were joined quickly by others. In the 1980s, Larry Hedges and Ingram Olkin (1985) provided the rigorous statistical proofs that established meta-analysis as an independent specialty within the statistical sciences.

Meta-analysis was not without its critics, and some criticisms persist. The value of quantitative synthesizing was questioned along lines similar to criticisms of primary data analysis. That is, some wondered whether using numbers to summarize research literatures might lead to the impression of the greater precision of results than was truly warranted by the data. However, much of the criticism stemmed less from issues in meta-analysis than from more general inappropriate synthesizing procedures, such as a lack of operational detail, which

were erroneously thought to be by-products of the use of quantitative procedures.

THE ELEMENTS OF META-ANALYSIS

Suppose a research synthesist is interested in whether fear-arousing advertisements can be used to persuade adolescents that smoking is bad. Suppose further that the (hypothetical) synthesist is able to locate eight studies, each of which examined the question of interest. Of these, six studies reported nonsignificant differences between attitudes of adolescents exposed and not exposed to fear-arousing ads, and two reported significant differences indicating less favorable attitudes held by adolescent viewers. One was significant at $p < .05$ and one at $p < .02$ (both two-tailed). Can the synthesist reject the NULL HYPOTHESIS that the ads had no effect?

The research synthesist could employ multiple methods to answer this question. First, the synthesist could cull through the eight reports, isolate those studies that present results counter to his or her position, discard these disconfirming studies due to methodological limitations, and present the remaining supportive studies as presenting the truth of the matter. Such a research synthesis would be viewed with extreme skepticism. It would contribute little to answering the question.

The Vote Count

As an alternative procedure, the synthesist could take each report and place it into one of three piles: statistically significant findings that indicate the ads were effective, statistically significant findings that indicate the ads created more positive attitudes toward smoking (in this case, this pile would have no studies), and nonsignificant findings that do not permit rejection of the hypothesis that the fear-arousing ads had no effect. The synthesist then would declare the largest pile the winner. In our example, the null hypothesis wins.

This vote count of significant findings has much intuitive appeal and has been used quite often. However, the strategy is unacceptably conservative. The problem is that chance alone should produce only about 5% of all reports falsely indicating that viewing the ads created more negative attitudes toward smoking. Therefore, depending on the number of studies, 10% or less of the positive and statistically significant findings might indicate a real difference due to the ads.

However, the vote-counting strategy requires that a minimum 34% of findings be positive and statistically significant before the hypothesis is ruled a winner. Thus, the vote counting of significant findings could and often does lead to the suggested abandonment of hypotheses (and effective treatment programs) when, in fact, no such conclusion is warranted.

A different way to perform vote counts in research synthesis involves (a) counting the number of positive and negative results, regardless of significance, and (b) applying the sign test to determine if one direction appears in the literature more often than would be expected by chance. This vote-counting method has the advantage of using all studies but suffers because it does not weight a study's contribution by its sample size. Thus, a study with 100 participants is given weight equal to a study with 1,000 participants. Furthermore, the revealed magnitude of the hypothesized relation (or impact of the treatment under evaluation) in each study is not considered—a study showing a small positive attitude change is given equal weight to a study showing a large negative attitude change. Still, the vote count of directional findings can be an informative complement to other meta-analytic procedures.

Combining Probabilities Across Studies

Taking these shortcomings into account, the research synthesist might next consider combining the precise probabilities associated with the results of each study. These methods require that the statistical tests (a) relate to the same hypothesis, (b) are independent, and (c) meet the initial statistical ASSUMPTIONS made by the primary researchers.

The most frequently applied method is called the method of adding Zs. In its simplest form, the adding Zs method involves summing Z-scores associated with p levels and dividing the sum by the square root of the number of inference tests. In the example of the eight studies of fear-arousing ads, the cumulative Z-score is 1.69, $p < .05$ (one-tailed). Thus, the null hypothesis is rejected.

Not all findings have an equal likelihood of being retrieved by the synthesist. Nonsignificant results are less likely to be retrieved than significant ones. This fact implies that the adding Zs method may produce a probability level that underestimates the chance of a Type I error. Rosenthal (1979) referred to this as the FILE DRAWER PROBLEM and wrote, "The extreme view of this problem . . . is that the journals are filled with

the 5% of studies that show Type I errors, while the file drawers back in the lab are filled with the 95% of studies that show insignificant (e.g., $p < .05$) results" (p. 638). The problem is probably not this dramatic, but it does exist.

The combining probabilities procedure overcomes the improper weighting problems of the vote count. However, it has severe limitations of its own. First, whereas the vote-counting procedure is overly conservative, the combining probabilities procedure is extremely powerful. In fact, it is so powerful that for hypotheses or treatments that have generated a large number of tests, rejecting the null hypothesis is so likely that it becomes a rather uninformative exercise. Furthermore, the combined probability addresses the question of whether an effect exists; it gives no information on whether that effect is large or small, important or trivial. Therefore, answering the question, "Do fear-arousing ads create more negative attitudes toward smoking in adolescents, yes or no?" is often not the question of greatest importance. Instead, the question should be, "How much of an effect do fear-arousing ads have?" The answer might be zero, or it might be either a positive or negative value. Furthermore, the research synthesist would want to ask, "What factors influence the effect of fear-arousing ads?" He or she realizes that the answer to this question could help make sound recommendations about how to both improve and target the smoking interventions. Given these new questions, the synthesists would turn to the calculation of average EFFECT SIZES, much like Pearson's correlation coefficient.

Estimating Effect Sizes

Cohen (1988) defined an effect size as "the degree to which the phenomenon is present in the population, or the degree to which the null hypothesis is false" (pp. 9–10). In meta-analysis, effect sizes are (a) calculated for the outcomes of studies (or sometimes comparisons within studies), (b) averaged across studies to estimate general magnitudes of effect, and (c) compared between studies to discover if variations in study outcomes exist and, if so, what features of studies might account for them.

Although numerous estimates of effect size are available, two dominate the social science literature. The first, called the d-index, is a scale-free measure of the separation between two group means. Calculating the d-index for any study involves dividing the

difference between the two group means by either their average STANDARD DEVIATION or the standard deviation of the control group. The second effect size metric is the *r*-index or correlation coefficient. It measures the degree of linear relation between two variables. In the advertising example, the *d*-index would be the appropriate effect size metric.

Meta-analytic procedures call for the weighting of effect sizes when they are averaged across studies. The reason is that studies yielding more precise estimates—that is, those with larger sample sizes—should contribute more to the combined result. *The Handbook of Research Synthesis* (Cooper & Hedges, 1994) provides procedures for calculating the appropriate weights. Also, this book describes how CONFIDENCE INTERVALS for effect sizes can be estimated. The confidence intervals can be used instead of the adding *Z*-scores method to test the null hypothesis that the difference between two means, or the size of a correlation, is zero.

Examining Variance in Effect Sizes

Another advantage of performing a meta-analysis of research is that it allows the synthesist to test hypotheses about why the outcomes of studies differ. To continue with the fear-arousing ad example, the synthesist might calculate average *d*-indexes for subsets of studies, deciding that he or she wants different estimates based on certain characteristics of the data. For example, the reviewer might want to compare separate estimates for studies that use different outcomes, distinguishing between those that measured likelihood of smoking versus attitude toward smoking. The reviewer might wish to compare the average effect sizes for different media formats, distinguishing print from video advertisements. The reviewer might want to look at whether advertisements are differentially effective for different types of adolescents, say, males and females.

After calculating the average effect sizes for different subgroups of studies, the meta-analyst can statistically test whether these factors are reliably associated with different magnitudes of effect. This is sometimes called a homogeneity analysis. Without the aid of statistics, the reviewer simply examines the difference in outcomes across studies, groups them informally by study features, and decides (based on unspecified inference rules) whether the feature is a significant predictor of variation in outcomes. At best, this method is imprecise. At worst, it leads to incorrect inferences.

In contrast, meta-analysis provides a formal means for testing whether different features of studies explain variation in their outcomes. Because effect sizes are imprecise, they will vary somewhat even if they all estimate the same underlying population value. Homogeneity analysis allows the synthesist to test whether SAMPLING ERROR alone accounts for this variation or whether features of studies, samples, treatment designs, or outcome measures also play a role. If reliable differences do exist, the average effect sizes corresponding to these differences will take on added meaning.

Three statistical procedures for examining variation in effect sizes appear in the literature. The first approach applies standard inference tests, such as ANOVA or MULTIPLE REGRESSION. The effect size is used as the DEPENDENT VARIABLE, and study features are used as independent or predictor variables. This approach has been criticized based on the questionable tenability of the underlying assumptions.

The second approach compares the variation in obtained effect sizes with the variation expected due to sampling error. This approach involves calculating not only the observed variance in effects but also the expected variance, given that all observed effects are estimating the same population value. This approach also can include adjustments to effect sizes to account for methodological artifacts.

The third approach, called homogeneity analysis, was mentioned above. It provides the most complete guide to making inferences about a research literature. Analogous to ANOVA, studies are grouped by features, and the average effect sizes for groups can be tested for homogeneity. It is relatively simple to carry out a homogeneity analysis. The formulas for homogeneity analysis are described in Cooper (1998) and Cooper and Hedges (1994).

In sum, then, a meta-analysis might contain four separate sets of statistics: (a) a frequency analysis of positive and negative results, (b) combinations of the *p* levels of independent inference tests, (c) estimates of average effect sizes with confidence intervals, and (d) homogeneity analyses examining study features that might influence study outcomes. The need for combined probabilities diminishes as the body of literature grows or if the author provides confidence intervals around effect size estimates. In addition to these basic components, the techniques of meta-analysis have grown to include many descriptive and

diagnostic techniques unique to the task of integrating the results of separate studies.

—Harris M. Cooper

REFERENCES

- Cohen, J. (1988). *Statistical power analysis in the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.
- Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Research*, 5, 3–8.
- Glass, G. V., & Smith, M. L. (1978). Meta-analysis of research on the relationship of class size and achievement. *Educational Evaluation and Policy Analysis*, 1, 2–16.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, 86, 721–735.
- Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, 3, 1243–1246.
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86, 638–641.
- Rosenthal, R., & Rubin, D. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 3, 377–415.

META-ETHNOGRAPHY

Meta-ethnography is an approach to synthesizing qualitative studies, allowing interpretations to be made across a set of studies (Noblit & Hare, 1988). In contrast to quantitative META-ANALYSIS (Glass, 1977), meta-ethnography does not aggregate data across studies as this ignores the context-specific nature of qualitative research. Instead, meta-ethnography involves synthesizing the interpretations across qualitative studies in which context itself is also synthesized. The approach allows for the development of more interpretive literature reviews, critical examinations of multiple accounts of similar phenomena, systematic comparisons of case studies to allow interpretive cross-case conclusions, and more formal syntheses of ETHNOGRAPHIC studies. In all these uses, meta-ethnography is both inductive and interpretive. It begins with the studies and inductively derives a synthesis by translating ethnographic and/or qualitative accounts into

the terms of each other to inductively derive a new interpretation of the studies taken collectively.

A meta-ethnography is accomplished through seven phases. First, one starts by identifying an intellectual interest that can be addressed by synthesizing qualitative studies. Second, it is necessary to decide what is relevant to the intellectual interest by considering what might be learned and for which audiences. Third, after conducting an exhaustive search for relevant studies, one reads the studies in careful detail. Fourth, it is necessary to consider the relationships of the studies to one another by making lists of key metaphors, phrases, ideas, and concepts and their relations to one another. The relationships may be reciprocal (similarity between studies is evident), refutational (studies oppose each other), or line of argument (studies of parts infer a whole). Fifth, the studies are translated into one another in the form of an analogy that one account is like the other except in some ways. Sixth, the translations are synthesized into a new interpretation that accounts for the interpretations in the different studies. Seventh, the synthesis is expressed or represented in some form appropriate to the audience of the synthesis: text, play, video, art, and so on.

For example, Derek Freeman’s refutation of Margaret Mead’s ethnography of Samoa was covered in the popular media in the 1980s. Freeman’s (1983) account was an attempt to discredit the work of Mead (1928), but the refutational synthesis in Noblit and Hare (1988) reveals that the studies come from different time periods, different paradigms of anthropology, different genders, and even different Samoas (American vs. Western). In short, the studies do not refute each other. They teach about different aspects of Samoan culture. Freeman’s claim that he refutes Mead and Mead’s claim to refute biological determinism both fail. Meta-ethnography, then, is a powerful tool used to reconsider what ethnographies are to mean to us taken together.

—George W. Noblit

REFERENCES

- Freeman, D. (1983). *Margaret Mead and Samoa*. Cambridge, MA: Harvard University Press.
- Glass, G. V. (1977). Integrating findings. *Review of Research in Education*, 5, 351–379.
- Mead, M. (1928). *Coming of age in Samoa*. New York: William Morrow.

Noblit, G. W., & Hare, R. D. (1988). *Meta-ethnography*. Newbury Park, CA: Sage.

METAPHORS

Metaphor refers to a figure of speech in which one thing is understood or described in terms of another thing. The two things are not assumed to be literally the same. For instance, one might talk metaphorically about a thinker “shedding light on an issue.” This does not mean that the thinker has actually lit a candle or turned on an electric light. In the social sciences, the use of metaphors has been studied as a topic, but also metaphors have contributed greatly to the development of theoretical thinking.

George Lakoff and Mark Johnson (1980), in their influential book *Metaphors We Live By*, argued that metaphors were socially and psychologically important. They suggested that culturally available metaphors influence how people think about issues. A metaphor that was once strikingly fresh becomes a shared cliché through constant repetition. The shared metaphor can provide a cognitive framework that directs our understanding of the social world. For instance, the competitiveness of politics, or of academic debate, is reinforced by the habitual use of warfare metaphors, which describe views as positions as being “attacked,” “defended,” or “shot down.”

Within the social sciences, metaphors can constitute the core of theoretical insights. For example, the metaphor of social life as theater has been enormously influential in the history of sociology. This dramaturgical metaphor has given rise to terms such as *social actor*, *social role*, and *social scene*. One can see this metaphor being used in novel, creative ways in the work of Erving Goffman, particularly in his classic book *The Presentation of Self in Everyday Life* (Goffman, 1959).

Metaphors, like coins, lose their shine through circulation. The idea of a “social role” might have once been metaphorical, but now that it has become a standard sociological term, its meaning tends to be understood literally. Role theorists assume that social roles actually exist and that they can be directly observed, even measured. The more recent metaphor—that the social world is a text or discourse to be read—has also traveled along a similar route. What was once novel and metaphorical, as in the work of Roland Barthes

(1977), becomes through repetition to be understood much more literally. The development of methodological practices, whether to identify social roles or to “read cultural practices,” hastens this journey from the excitingly metaphorical to the routinely literal.

It is too simple to claim that metaphors in the social sciences invariably start by being nonliteral, for some have always been inherently ambiguous. Psychologists have often used the metaphor of “the mind as machine.” Even early theorists seem to have been uncertain whether the mind should be understood to be “like a machine” or whether it was actually a machine (Soyland, 1994). What is clear, however, is that as later scientists take up the idea, transforming it into research programs with established methodologies, so progressively the metaphorical element gives way to the literal. When this happens, a field of inquiry may become successfully established, but it will be vulnerable to assault by a younger generation of thinkers, who are inspired by new metaphorical insights.

—Michael Billig

REFERENCES

- Barthes, R. (1977). *Image music text*. London: Fontana.
 Goffman, E. (1959). *The presentation of self in everyday life*. New York: Anchor.
 Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
 Soyland, A. J. (1994). *Psychology as metaphor*. London: Sage.

METHOD VARIANCE

Method variance refers to the effects the method of measurement has on the things being measured. According to classical test theory, the observed VARIANCE among participants in a study on a variable can be attributable to the underlying TRUE SCORE or construct of interest plus RANDOM ERROR. Method variance is an additional source of variance attributable to the method of assessment. Campbell and Fiske (1959) gave examples of apparatus effects with Skinner boxes used to condition rats and format effects with psychological scales.

It has been widely assumed in the absence of solid evidence that when the same method is used to assess different variables, method effects (also called monomethod bias) will produce a certain level of SPURIOUS covariation among those variables. Many researchers,

for example, are suspicious of survey studies in which all variables are contained in the same questionnaire, assuming that by being in the same questionnaire, there is a shared method that produces spurious correlation among items. Research on method variance has cast doubt on such assumptions. For example, Spector (1987) was unable to find evidence to support that method effects are widespread in organizational research. One of his arguments was that if method variance produced spurious correlations, why were so many variables in questionnaire studies unrelated? Williams and Brown (1994) conducted analyses showing that in most cases, the existence of variance due to method would serve to attenuate rather than inflate relations among variables.

Spector and Brannick (1995) noted that the traditional view that a method (e.g., questionnaire) would produce method effects across all variables was an oversimplification. They discussed how constructs and methods interacted, so that certain methods used to assess certain variables might produce common variance. Furthermore, they suggested that it was not methods themselves but rather features of methods that were important. For example, SOCIAL DESIRABILITY is a well-known potential BIAS of self-reports about issues that are personally sensitive or threatening (e.g., psychological adjustment), but it does not bias reports of nonsensitive issues. If the method is a self-report and the construct is sensitive, a certain amount of variance will be attributable to social desirability. With such constructs, it may be necessary to find a procedure that can control these effects.

—Paul E. Spector

REFERENCES

- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Spector, P. E. (1987). Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem? *Journal of Applied Psychology*, 72, 438–443.
- Spector, P. E., & Brannick, M. T. (1995). The nature and effects of method variance in organizational research. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology: 1995* (pp. 249–274). West Sussex, England: Wiley.
- Williams, L. J., & Brown, B. K. (1994). Method variance in organizational behavior and human resources research: Effects on correlations, path coefficients, and hypothesis testing. *Organizational Behavior and Human Decision Processes*, 57, 185–209.

METHODOLOGICAL HOLISM

The opposite of METHODOLOGICAL INDIVIDUALISM, this doctrine holds that social wholes are more than the sum of individual attitudes, beliefs, and actions and that the whole can often determine the characteristics of individuals. Holism has been prominent in philosophy and social science since Hegel, and, arguably, it has its roots in the writings of Plato. Methodological holism (often abbreviated to *holism*) takes a number of forms across social science disciplines. Although very different in their views and emphases, Marx, Dewey, Durkheim, and Parsons can all be regarded as holists.

The quintessential holist thinker was the sociologist Emile Durkheim (1896/1952), who argued that social effects must be explained by social causes and that psychological explanations of the social were erroneous. His evidence for this came from his study of suicide statistics, over time, in several European countries. Although suicide rates rose and fell, the proportions attributed to individual causes by coroners remained constant, suggesting that even such an individual act as suicide was shaped by social forces.

Holism in anthropology and sociology was much encouraged by the biological “organicism” of Darwinism. Herbert Spencer (Dickens, 2000, pp. 20–24), for example, saw society evolving through the adaptation of structures, in much the same way as biological structures evolved. A more sophisticated version of organic thinking was advanced by Talcott Parsons (1968), in which, like the organs of the human body, the constituent parts of society function to maintain each other. Parsons’s holism was explicitly functionalist, but critics of holism and functionalism have maintained that the latter is necessarily at least implicit in the former.

Notwithstanding Popper’s and Hayek’s political criticisms of holism (see METHODOLOGICAL INDIVIDUALISM), methodological critique has centered on the incompleteness of explanations that do not make ultimate reference to individual beliefs, desires, and actions. For example, although a reference to the Treaty of Versailles and the subsequent war reparations required of Germany might be offered as historical explanations of the origins of World War II, this cannot stand on its own without reference to the views and actions of agents responsible for its enactment and the response to it by other individuals, particularly Hitler.

Durkheim, as well as many after him, believed that sociology particularly depended, as a scientific discipline, on the existence of social facts with an ontological equivalence to that of objects in the physical world. In recent years, however, the language of (and the debate about) methodological holism and individualism has been superseded by that of “structure and agency.” Only a few thinkers now refer to themselves as “holists” or “individualists,” and although extreme forms of each position can be logically demonstrated, much recent debate concerns the extent to which the social can be explained by individual agency, or agents’ beliefs, desires, and actions by the social.

In recent years, sociologists such as Anthony Giddens (1984) and philosophers such as Roy Bhaskar (1998) have attempted to bridge the gap between holism and individualism by advancing versions of “structuration theory,” which, in Giddens’s case, focuses on social practices, which he argues produce and are produced by structure (Giddens, 1984).

—Malcolm Williams

REFERENCES

- Bhaskar, R. (1998). *The possibility of naturalism* (2nd ed.). London: Routledge Kegan Paul.
- Dickens, P. (2000). *Social Darwinism*. Buckingham, UK: Open University Press.
- Durkheim, E. (1952). *Suicide*. London: Routledge Kegan Paul. (Original work published 1896)
- Giddens, A. (1984). *The constitution of society*. Cambridge, UK: Polity.
- Parsons, T. (1968). *The structure of social action* (3rd ed.). New York: Free Press.

METHODOLOGICAL INDIVIDUALISM

There is a long tradition in the social sciences—going back to Hobbes and continued through J. S. Mill, Weber, and, more recently, George Homans and J. W. N. Watkins—that insists that explanations of social phenomena must derive from the characteristics and dispositions of individual agents. In recent decades, this *methodological* approach has become somewhat conflated with individualist political and economic doctrines (Lukes, 1994).

Most of the arguments for methodological individualism have been couched in response to the

opposite doctrine, METHODOLOGICAL HOLISM (and vice versa). Individualists reject the latter’s claim that there can be social wholes existing beyond or separate to individuals—not by saying these social wholes do not exist, but rather claiming that the social world is brought into being by individuals acting according to the “logic of the situation.” That is, individuals create social situations by directing their actions according to the expected behavior of other individuals in particular situations. Complex social situations (e.g., economic systems) are said to result from particular alignments of individual dispositions, beliefs, and physical resources or characteristics. In explaining a given economic system, reference to constituent parts (e.g., the role of merchant banks) can only be seen as partial explanation, although they may be constructed as heuristic models. A full explanation must, at rock bottom, refer to actual or possibly “ideal” (statistical) agents.

Since J. S. Mill, there has been an association between individualist political and economic philosophy in liberalism and an individualist methodological approach. This association became more explicit in the writings of Frederick von Hayek and Karl Popper. Popper maintained that “collective” (holist) explanation blinded us to the truth of the matter that the functioning of all social institutions results from the attitudes, decisions, and actions of individuals. Moreover, he claimed, collectivist views about what *is* lead to collectivist views about what *should* be and, ultimately, the totalitarianism of fascism and communism (Popper, 1986).

Whatever the merits or demerits of this political argument, it is widely acknowledged to be a non sequitur from the methodological argument, which has been associated with thinkers of both the right and left (Phillips, 1976). There are, however, a number of methodological objections to the doctrine, foremost among which is that it leads to the conclusion that social characteristics can only be described in terms of individual psychologies, yet most individual characteristics (such as being a banker, a teacher, or a police officer) can only be explained in terms of social institutions. A second *REALIST* objection concerns the ontological status of social practices or objects. Methodological individualists maintain that, unlike physical objects that have real properties, social phenomena are simply collective mental constructions. The response to this is that many things in the physical world can only be known by their observed effects (e.g., gravity), and much the same is true of social

phenomena such as a foreign policy or a body of law, but they, nonetheless, are real for this.

—Malcolm Williams

REFERENCES

- Lukes, S. (1994). Methodological individualism reconsidered. In M. Martin & L. McIntyre (Eds.), *Readings in the philosophy of social science* (pp. 451–459). Cambridge: MIT Press.
- Phillips, D. (1976). *Holistic thought in social science*. Stanford, CA: Stanford University Press.
- Popper, K. (1986). *The poverty of historicism*. London: ARK.

METRIC VARIABLE

A metric variable is a variable measured QUANTITATIVELY. The distance between the values is equal (e.g., the distance from scores 4 to 6 is the same as the distance from scores 6 to 8). INTERVAL, RATIO, and CONTINUOUS variables can have metric measurement and be treated as metric variables.

—Michael S. Lewis-Beck

See also INTERVAL

MICRO

Micro is the opposite of macro. Microanalysis is small scale, or at the individual level. Macroanalysis is large scale, or at the aggregate level.

—Michael S. Lewis-Beck

MICROSIMULATION

As with SIMULATION, the major purpose of microsimulation is to generate human, social behavior using computer ALGORITHMS. The distinctive feature of microsimulation, however, is that such a MODEL operates at the MICRO level, such as a person, family or household, or company. By simulating large representative MACRO-level POPULATIONS of these micro-level entities, conclusions are drawn to be applicable to an entire society or country. Microsimulation can

play a role not only in BASIC RESEARCH but also in assisting policymaking, such as helping U.S. states weigh different approaches to welfare reform.

—Tim Futing Liao

MIDDLE-RANGE THEORY

An idea that is associated with Robert Merton and his *Social Theory and Social Structure* (1949, 1957, 1968), middle-range theories fall between the working hypotheses of everyday research and unified, general theories of social systems. Merton was critical of the sociological quest for general theories, exemplified by Talcott Parsons's project, because he believed such theories to be too abstract to be empirically testable. However, he did not regard empirical generalizations about the relationship between variables as an adequate alternative. He believed, instead, that sociology should be committed to developing middle-range theories that are both close enough to observed data to be testable and abstract enough to inform systematic theory development. These theories are based on uniformities found in social life and take the form of causal laws (as opposed to empirical description). Merton meant for theories of the middle range to answer his own call for empirically informed sociological theorizing. His hope was that, over time, this program would generate an empirically verified body of social theory that could be consolidated into a scientific foundation for sociology.

Merton's commitments have roots in the philosophy of science and the empirically oriented sociology of his day. A sensitivity to the intellectual limits of extreme generality and particularity can be traced from Plato's *Theaetetus* to Francis Bacon and J. S. Mill. In *A System of Logic*, Mill (1865/1973) notably advocated limited causal theories based on classes of communities. Among sociologists, Merton cites Emile Durkheim's (1951) *Suicide*, with its limited causal theories based on national cases, as the classic example of middle-range theorizing. The ideal grew popular among American sociologists oriented toward research and concerned with the empirical applicability of theory during the first half of the past century. As it did so, space was created for theoretically informed research of a more scientific character than the grand theories of late 19th-century sociology.

Modeled on scientific inquiry, middle-range theories characteristically posit a central idea from which specific hypotheses can be derived and tested. Middle-range theorizing begins with a concept and its associated imagery and develops into a set of associated theoretical puzzles. For Merton, these puzzles centered on the operation of the social mechanisms that give rise to social behavior, organization, and change. Middle-range theorizing should produce causal models of social processes (such as theory of reference groups or social differentiation). The designation of the specific conditions under which a theory applies (scope conditions) fixes it in the middle range. Theorizing of this sort can be fruitfully conducted at both the micro-behavioral (such as the social psychology of small groups) and macro-structural levels (such as the interrelation among elements of social structure) as well as between the two (micro-macro or macro-micro relations).

Although structural functionalism is no longer *de mode*, sociological research in the spirit of middle-range theory continues to be produced, most notably by exchange and network theorists. The ideal of middle-range theory has again been the subject of attention and debate over the past few years. A collection of essays edited by Peter Hedstrom and Richard Swedberg (1998), featuring contributions by academicians from diverse backgrounds, is based on the premise that middle-range theorizing is the key to developing causal models of social processes, specifically focusing on the micro-macro linkage. In a similar spirit to Merton's program, these scholars call for empirically confirmed, rigorously delimited causal sociological theorizing.

—George Ritzer and Todd Stillman

REFERENCES

- Durkheim, E. (1951). *Suicide*. New York: Free Press.
- Hedstrom, P., & Swedberg, R. (1998). *Social mechanisms: An analytical approach to social theory*. Cambridge, UK: Cambridge University Press.
- Merton, R. (1949). *Social theory and social structure*. New York: Free Press.
- Merton, R. (1968). *Social theory and social structure* (3rd ed.). New York: Free Press.
- Mill, J. S. (1973). *A system of logic*. Toronto: University of Toronto Press. (Original work published 1865)
- Sztompka, P. (1990). *Robert Merton: An intellectual profile*. London: Macmillan.

MILGRAM EXPERIMENTS

In Milgram's (1963, 1974) obedience experiments, research participants, in the role of "teacher," were instructed to administer increasingly severe electric shocks to another participant, the "learner," for errors in a learning task. An unexpectedly high percentage of participants followed the experimenter's instructions to administer the maximum (apparently extremely painful) shock level. In fact, no shocks were administered, and all parties to the experiment, with the exception of the "teacher," were members of Milgram's research team. Despite their being, in many respects, so *unlike* any other research, these studies have been, for almost 40 years, the key reference point in discussions of ethical issues in social science research. Why have they been so prominent in the specific context of ethical issues?

1. The Milgram experiments have become the best-known and, for some, the most important research in psychology. Observing high rates of destructive obedience on the part of normal persons, they have been interpreted as revealing a major cause of the Nazi Holocaust (Miller, 1995). The unusual circumstance of a research program being both extraordinarily famous but also extremely controversial ethically has contributed to the unprecedented impact of these studies and created a highly instructive benchmark for discussing research ethics *in general*.

2. Milgram's first published account (1963) presented highly graphic portraits of stress and tension experienced by the research participants: "I observed a mature and initially poised businessman enter the laboratory smiling and confident. Within 20 minutes he was reduced to a twitching, stuttering wreck, who was rapidly approaching a point of nervous collapse" (Milgram, 1963, p. 377). It was in the context of these reports that the Milgram experiments came to be viewed as exposing participants to possibly dangerous levels of psychological risk.

3. A landmark citation was an early ethical critique of the Milgram experiments by Diana Baumrind (1964), published in the *American Psychologist*. Baumrind accused Milgram of (a) violating the experimenter's obligation to protect the emotional welfare of participants, (b) conducting an inadequate postexperimental DEBRIEFING, and (c) using a flawed methodology that produced highly questionable

results, thus arguing that the benefits of Milgram's research were not worth the ethical costs. She concluded that Milgram's participants would experience a loss of trust in authority figures and a permanent loss of self-esteem. Although Baumrind's criticisms were subsequently rebutted by Milgram and many others (Miller, 1986), her commentary had a powerful impact and became a model for ethical criticism.

4. A key feature of the Milgram experiments was the coercion built into the method. Participants who expressed a desire to withdraw from the study, were continuously prodded by the experimenter to continue in their role of shocking the other participant. By current ethical standards of the American Psychological Association (APA), this coercive aspect could make the obedience research unethical:

APA Ethics Codes Revision Draft 6,
October 21, 2001

8.02 INFORMED CONSENT to Research

When obtaining informed consent as required in Standard 3.10, Informed Consent, psychologists inform participants about (1) the purpose of the research, expected duration, and procedures; (2) their right to decline to participate and to withdraw from the research once participation has begun.

Although the use of DECEPTION and a lack of a truthful informed consent were also important features of the Milgram experiments, these factors, in themselves, are not a crucial feature of the legacy of the Milgram experiments. A significant proportion of contemporary psychological research (including the Ethical Guidelines of the APA) endorses a judicious use of deception (Epley & Huff, 1998).

In conclusion, the Milgram experiments have had a profound impact on considerations of ethical issues in research. The invaluable, often impassioned, debates surrounding these experiments have sensitized generations of students and researchers to the emotional and cognitive risk factors in research methods. The routine use of institutional review boards to assess, independent of primary investigators, ethical aspects of research prior to the conduct of such inquiry can also be viewed as a highly significant legacy of the Milgram experiments. Finally, the Milgram experiments have taught us that there may be inevitable ethical costs in the conduct of highly significant and potentially beneficial

psychological research. The rights of researchers to inquire must be honored while simultaneously gauging and protecting the vulnerabilities of their participants.

—Arthur G. Miller

REFERENCES

- Baumrind, D. (1964). Some thoughts on ethics after reading Milgram's "Behavioral study of obedience." *American Psychologist*, 19, 421–423.
- Epley, N., & Huff, C. (1998). Suspicion, affective response, and educational benefit of deception in psychology research. *Personality and Social Psychology Bulletin*, 24, 759–768.
- Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.
- Milgram, S. (1974). *Obedience to authority: An experimental view*. New York: Harper & Row.
- Miller, A. G. (1986). *The obedience experiments: A case study of controversy in social science*. New York: Praeger.
- Miller, A. G. (1995). Constructions of the obedience experiments: A focus upon domains of relevance. *Journal of Social Issues*, 51, 33–53.

MISSING DATA

In the social sciences, nearly all QUANTITATIVE DATA sets have missing data, which simply means that data are missing on some variables for some cases in the data set. For example, annual income may be available for 70% of the respondents in a sample but missing for the other 30%. There are many possible reasons why data are missing. Respondents may refuse to answer a question or may accidentally skip it. Interviewers may forget to ask a question or may not record the response. In longitudinal studies, persons may die or move away before the next follow-up interview. If data sets are created by merging different files, there may be matching failures or lost records.

Whatever the reasons, missing data typically create substantial problems for statistical analysis. Virtually all standard statistical methods presume that every case has complete information on all variables of interest. Until recently, there has been very little guidance for researchers on what to do about missing data. Instead, researchers have chosen from a collection of ad hoc methods that had little to recommend them except long and widespread usage. Fortunately, two principled methods for handling missing data are now available: maximum likelihood and multiple imputation. If appropriate assumptions are met, these methods

have excellent statistical properties and are now within the reach of the average researcher. This entry first reviews conventional methods for handling missing data and then considers the two newer methods. To say meaningful things about the properties of any of these methods, one must first specify the assumptions under which they are valid.

ASSUMPTIONS

To keep things simple, let us suppose that a data set contains only two variables, X and Y . We observe X for all cases, but data are missing on Y for, say, 20% of the cases. We say that data on Y are *missing completely at random* (MCAR) if the probability that data are missing on Y depends on neither Y nor X . Formally, we have $\Pr(Y \text{ is missing} | X, Y) = \Pr(Y \text{ is missing})$. Many missing data techniques are valid only under this rather strong assumption.

A weaker assumption is that the data are *missing at random* (MAR). This assumption states that the probability that data are missing on Y may depend on the value of X but does *not* depend on the value of Y , holding X constant. Formally, we have $\Pr(Y \text{ is missing} | X, Y) = \Pr(Y \text{ is missing} | X)$. Suppose, for example, that Y is a measure of income and X is years of schooling. The MAR assumption would be satisfied if the probability that income is missing depends on years of schooling, but within each level of schooling, the probability of missing income does not depend on income itself. Clearly, if the data are missing completely at random, they are also missing at random.

It is certainly possible to test whether the data are missing completely at random. For example, one could compare males and females to see whether they differ in the proportion of cases with missing data on particular variables. Any such difference would be a violation of MCAR. However, it is impossible to test whether the data are missing at random. For obvious reasons, you cannot tell whether persons with high income are less likely than those with moderate income to have missing data on income.

Missing data are said to be *ignorable* if the data are MAR and, in addition, the parameters governing the missing data mechanism are completely distinct from the parameters of the model to be estimated. This somewhat technical condition is unlikely to be violated in the real world. Even if it were, methods that assume ignorability would still perform very well if the data

are only MAR. It is just that one could do even better by modeling the missing data mechanism as part of the estimation process. So in practice, *missing at random* and *ignorability* are often used interchangeably.

CONVENTIONAL METHODS

The most common method for handling missing data—and the one that is the default in virtually all statistical packages—is known as LISTWISE DELETION or *complete case analysis*. The rule for this method is very simple: Delete any case that has missing data on any of the variables in the analysis of interest. Because the cases that are left all have complete data, statistical methods can then be applied with no difficulty.

Listwise deletion has several attractive features. Besides being easy to implement, it has the virtue that it works for *any* statistical method. Furthermore, if the data are missing completely at random, listwise deletion will not introduce any bias into the analysis. That is because under MCAR, the subset of the sample with complete data is effectively a SIMPLE RANDOM SAMPLE from the original data set. For the same reason, STANDARD ERRORS estimated by conventional methods will be valid estimates of the true standard errors.

The big problem with listwise deletion is that it often results in the elimination of a large fraction of the cases in the data set. Suppose, for example, that the goal is to estimate a multiple regression model with 10 variables. Suppose, furthermore, that each variable has 5% missing data and that the probability of missing on any variable is unrelated to whether or not data are missing on any other variables. Then, on average, half the cases would be lost under listwise deletion. The consequence would be greatly increased standard errors and reduced statistical power.

Other conventional methods are designed to salvage some of the data lost by listwise deletion, but these methods often do more harm than good. PAIRWISE DELETION, also known as *available case analysis*, is a method that can be easily implemented for a wide class of linear models, including MULTIPLE REGRESSION and FACTOR ANALYSIS. For each pair of variables in the analysis, the correlation (or covariance) is estimated using all the cases with no missing data for those two variables. These correlations are combined into a single correlation matrix, which is then used as input to linear modeling programs. Under MCAR, estimates produced by this method are consistent (and thus approximately unbiased), but standard errors and associated

test statistics are typically biased. Furthermore, the method may break down entirely (if the correlation matrix is not positive definite).

Dummy variable adjustment is a popular method for handling missing data on independent variables in a regression analysis. When a variable X has missing data, a dummy (indicator) variable D is created with a value of 1 if data are present on that variable and 0 otherwise; for cases with missing data, X is set to some constant (e.g., the mean for nonmissing cases). Both X and D are then included as independent variables in the model. Although this approach would seem to incorporate all the available information into the model, it has been shown to produce biased estimates even if the data are MCAR.

A variety of missing data methods fall under the general heading of *IMPUTATION*, which means to substitute some plausible values for the missing data. After all the missing data have been imputed, the statistical analysis is conducted as if there were no missing data. One of the simplest methods is *unconditional mean imputation*: If a variable has any missing data, the mean for the cases without missing data is substituted for those cases with missing data. But this method is known to produce biased estimates of many parameters; in particular, variances are always underestimated. Somewhat better is *conditional mean imputation*: If a variable has missing data, one estimates a linear regression of that variable on other variables for those cases without missing data. Then, the estimated regression line is used to generate predicted values for the missing data. This method can produce approximately unbiased estimates of regression coefficients if the data are MCAR.

Even if one can avoid bias in parameter estimates, however, all conventional methods of imputation lead to underestimates of standard errors. The reason is quite simple. Standard methods of analysis presume that all the data are real. If some data are imputed, the imputation process introduces additional sampling variability that is not adequately accounted for.

MAXIMUM LIKELIHOOD

Much better than conventional methods is *MAXIMUM LIKELIHOOD (ML)*. Under a correctly specified model, ML can produce estimates that are consistent (and, hence, approximately unbiased), *ASYMPTOTICALLY* efficient, and asymptotically normal, even when data are missing (Little & Rubin, 2002). If the

missing data mechanism is ignorable, constructing the likelihood function is relatively easy: One simply sums or integrates the conventional likelihood over the missing values. Maximization of that likelihood can then be accomplished by conventional numerical methods.

Although the general principles of ML for missing data have been known for many years, only recently have practical computational methods become widely available. One very general numerical method for getting ML estimates when data are missing is the EM algorithm. However, most EM software only produces estimates of means, variances, and covariances, under the assumption of multivariate normality. Furthermore, the method does not produce estimates of standard errors.

Much more useful is the incorporation of ML missing data methods into popular structural equations modeling programs, notably LISREL (www.ssicentral.com/lisrel), AMOS (www.smallwaters.com), and MPLUS (www.statmodel.com). Under the assumption of multivariate normality, these programs produce ML estimates for a wide class of linear models, including linear regression and factor analysis. Estimated standard errors make appropriate adjustments for the missing data.

For categorical data analysis, ML methods for missing data have been implemented in the program *ℓEM* (http://cwis.kub.nl/~fsw_1/mto/mto3.htm). This software estimates a large class of LOG-LINEAR and LATENT CLASS models with missing data. It will also estimate LOGISTIC REGRESSION models when all predictor variables are discrete.

MULTIPLE IMPUTATION

Although ML is a very good method for handling missing data, it still requires specialized software to do the statistical analysis. At present, such software is only available for a rather limited class of problems. A much more general method is *multiple imputation (MI)*, which can be used for virtually any kind of data and for any kind of analysis. Furthermore, the analysis can be done with conventional software.

The basic idea of MI is simple. Instead of imputing a single value for each missing datum, multiple values are produced by introducing a random component into the imputation process. As few as three imputed values will suffice for many applications, although more are certainly preferable. These multiple imputations are used to construct multiple data sets. So if

three imputed values are generated for each missing value, the result will be three “completed” data sets, all with the same values for the nonmissing data but with slightly different values for the imputed data.

Next, the analysis is performed on each completed data set using conventional software. Finally, the results of these replicated analyses are combined according to some very simple rules. For parameter estimates, one takes the mean of the estimates across the replications. Standard errors are calculated by averaging the squared standard errors across replications, adding the variance of the parameter estimates across replications (with a small correction factor), and then taking the square root. The formula is

$$SE(\bar{b}) = \sqrt{\frac{1}{M} \sum_k s_k^2 + \left(1 + \frac{1}{M}\right) \left(\frac{1}{M-1}\right) \sum_k (b_k - \bar{b})^2},$$

where b_k is the parameter estimate in replication k , s_k is the estimated standard error of b_k , and M is the number of replications.

This method accomplishes three things. First, appropriate introduction of random variation into the imputation process can eliminate the biases in parameter estimation that are endemic to deterministic imputation methods. Second, repeating the process several times and averaging the parameter estimates results in more stable estimates (with less sampling variability). Finally, the variability in the parameter estimates across replications makes it possible to get good estimates of the standard errors that fully reflect the uncertainty in the imputation process.

How do the statistical properties of MI compare with those of ML? If all the assumptions are met, MI produces estimates that are consistent and asymptotically normal, just like ML. However, they are not fully efficient unless there are an infinite number of imputations, something not likely to be seen in practice. In most cases, however, only a handful of imputations are necessary to achieve close to full efficiency.

Now that we have the basic structure of multiple imputation, the next and obvious question is how to generate the imputations. There are many different ways to do this. Like ML, the first step is to choose a model for the data. The most popular model, at present, is the multivariate normal model, which implies that (a) all variables are normally distributed and (b) each variable can be represented as a linear function of all

the other variables. The force of this assumption is that all imputations are based on linear regressions. The random variation in imputed values is accomplished by making random draws from NORMAL DISTRIBUTIONS that have the estimated residual variance from each linear regression. These random draws are added to the predicted values for each case with missing data.

The details of how to accomplish all this can get rather technical and are not worth pursuing here. (Schafer [1997] gives a very thorough treatment; Allison [2001] provides a less technical discussion.) Fortunately, there are now several readily available computer programs that produce multiple imputations with a minimum of effort on the part of the user. SAS has recently introduced procedures for producing multiple imputations and for combining results from multiple imputation analyses (www.sas.com). SOLAS is a commercial stand-alone program for doing MI (www.statsolusa.com). Freeware stand-alone programs include NORM (www.stat.psu.edu/~jls) and AMELIA (gking.harvard.edu/stats.shtml).

MI can seem very complicated and intimidating to the novice. Although it is certainly more work than most conventional methods for handling missing data, good software can make MI a relatively painless process. For example, to do MI for multiple regression in SAS, one needs only two lines of additional program code to do the imputation, one additional line in the regression procedure, and two lines of subsequent code to combine the results.

OTHER APPROACHES

Before closing, other methods currently under development deserve a brief mention. Although MI methods based on the multivariate normal model are the most widely and easily used, other parametric models may be less restrictive. For example, imputations can be generated under a log-linear model, a general location scale model (for both categorical and quantitative variables), or a sequential generalized regression model. There are also nonparametric and semiparametric methods available for multiple imputation.

Although the MI and ML methods discussed here are all based on the assumption that the missing data mechanism is ignorable, it is quite possible to apply both methods to nonignorable situations in which the probability of missing data on Y depends on the value of Y itself. Heckman’s method for handling selectivity

bias is a good example of how to accomplish this by ML. However, methods for nonignorable missing data tend to be extremely sensitive to the choice of models for the missing data mechanism.

—Paul D. Allison

REFERENCES

- Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.
 Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley.
 Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

MISSPECIFICATION

Among the assumptions employed in classical linear REGRESSION, it is taken that the researcher has specified the “true” model for estimation. Misspecification refers to several and varied conditions under which this assumption is not met. The consequences of model misspecification for PARAMETER ESTIMATES and STANDARD ERRORS may be either minor or serious depending on the type of misspecification and other mitigating circumstances. Tests for various kinds of misspecification lead us to consider the bases for choice among competing MODELS.

TYPES OF MISSPECIFICATION

Models may be incorrectly specified in one or more of three basic ways: (a) with respect to the set of REGRESSORS selected, (b) in the measurement of model variables, and (c) in the specification of the stochastic term or ERROR, u_i . First, estimation problems created by incorrect selection of regressors will vary depending on whether the researcher has (a) omitted an important variable from the model (see OMITTED VARIABLE), (b) included an extraneous variable, or (c) misspecified the functional form of the relationship between X_i and Y . Under most circumstances, the first condition—omission of a relevant variable from the model—results in the most serious consequences for model estimates. Specifically, if some variable Z_i is responsible for a portion of the variation in Y , yet Z_i is omitted from the model, then parameter estimates will be both biased and inconsistent, unless Z is perfectly independent of the included variables. The extent and direction of BIAS depend on the CORRELATION between the omitted

Z_i and the included variables X_k . Inclusion of an extraneous variable (i.e., “overfitting” the model) is generally less damaging because all model parameters remain unbiased and consistent. However, a degree of inefficiency in those estimates is introduced in that the calculated standard errors for each parameter take account of the correlation between each “true” variable and the extraneous X_k . Because that correlation is likely to be nonzero in the sample, estimated standard errors are inflated unnecessarily. Misspecification of the functional form linking Y to X can be seen as a mixed case of overfitting as well as underfitting. Imagine a “true” model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + u_i$. If we now estimate a model in which Y is a simple linear function of X_1 and X_2 (rather than the square of X_2 found in the true model)—that is, $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + e_i$ —then parameter estimates and standard errors for α_0 , α_1 , and α_2 will be vitiated by the inclusion of the extraneous X_2 as well as by the omission of the squared term X_2^2 .

Beyond the incorrect choice of regressors, measurement errors in the dependent and independent variables may be viewed as a form of model misspecification. This is true simply because such measurement error may result in the same variety of problems in parameter estimation as discussed above (i.e., estimated parameters that are biased, inconsistent, or inefficient). Finally, proper identification of the process creating the stochastic errors, u_i , can be seen as a model specification issue. To see this point, begin with the common assumption that the errors are distributed normally with mean zero and variance σ^2 (i.e., $u_i \sim N(0, \sigma^2)$). Now imagine that the true errors are generated by any other process (e.g., perhaps the errors may exhibit HETEROSKEDASTICITY or they are distributed log normal). In this instance, we will again face issues of bias, inconsistency, or inefficiency depending on the particular form of misspecification of the error term.

TESTS FOR MISSPECIFICATION

Good THEORY is by far the most important guide we have in avoiding misspecification. But given that social science theories are often weak or tentative, empirical tests of proper specification become more important. A variety of tests are associated with each of the specification problems described above. One should always begin with analysis of RESIDUALS from the initial estimation. Cases sharing common traits that cluster above or below the regression line often

suggest an omitted variable. General patterns in those residuals may suggest a nonlinear relationship between X and Y (i.e., an issue of functional form) or perhaps other pathologies such as heteroskedasticity. In short, analysis of residuals is an important means whereby the researcher comes to know the cases and their relationship to the model specified. Specification, especially when theory is weak, is an iterative process between model and data.

If one suspects that some omitted variable, Z_i , may be part of the true model, the well-known DURBIN-WATSON d statistic may be employed to test this hypothesis. Simply arrange the original ordinary least squares (OLS) residuals according to values of Z_i and compute d by the usual means. Rejection of the null hypothesis indicates that Z_i has indeed been incorrectly omitted from the original specification. More general tests for omitted variables include Ramsey's RESET and the Lagrange multiplier (LM) test. Neither RESET nor the LM test, however, is capable of identifying the missing variable(s). That important step can only be informed by refinement of theory and deepened understanding of the cases, perhaps through techniques such as analysis of residuals. Gujarati (2003, pp. 518–524) describes specific steps to be taken with these tests in greater detail.

MODEL SPECIFICATION AND MODEL SELECTION

Given that we can never be certain that we have correctly identified the “true” model, modern practitioners argue that our efforts are better focused on assessing the adequacy of contending models of a given phenomenon Y . Imagine two contending models of Y . If all explanatory variables in one contender are a perfect subset of variables included in a second model, then we may say the first is “nested” within the second. Recalling that the regressand must be identical across equations, well-known statistics such as adjusted R -SQUARED, the F test and the likelihood ratio test (for nonlinear estimators) are entirely appropriate to assess the empirical adequacy of one model against another. Competing models are “nonnested” when the set of regressors in one contender overlaps partially or not at all with the regressors included in the second contending model of Y . Keynesian versus monetarist models of gross national product (GNP) growth, for example, would be nonnested competitors. Comparing the relative performance of nonnested models, the

researcher may turn again to adjusted R^2 and certain other GOODNESS-OF-FIT MEASURES such as the Akaike Information Criterion. Clarke (2001, pp. 732–735) also describes application of the Cox test and Vuong test for these purposes. Finally, the activity of model selection lends itself quite directly to BAYESIAN INFERENCE, especially when, practically speaking, our data are not subject to repeated sampling. The subject of model selection from a Bayesian perspective is introduced in Leamer (1978) and updated in an applied setting in Clarke (2001).

—Brian M. Pollins

REFERENCES

- Clarke, K. A. (2001). Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science*, 45, 724–744.
- Gujarati, D. N. (2003). *Basic econometrics* (4th ed.). Boston: McGraw-Hill.
- Leamer, E. E. (1978). *Specification searches*. New York: John Wiley.

MIXED DESIGNS

Mixed designs have long been a mainstay of experimentation, but in recent years, their importance has been made greater still by the development of a new statistical methodology. This methodology, called *mixed modeling*, enhances the VALIDITY of analyses of the data generated by a mixed design.

Mixed design is a term that sometimes has been loosely applied to any research plan involving both QUALITATIVE and QUANTITATIVE variables or, alternatively, involving both manipulated and unmanipulated variables. However, psychologists usually define it as a formal research plan that includes both *between*-subjects and *within*-subjects factors, and statisticians generally define it as a RESEARCH DESIGN that involves the estimation of both *fixed* factor effects and *random* factor effects.

With a between-subjects factor, each experimental unit is exposed to only a single level of that factor. The factor levels are mutually exclusive and (assuming a balanced design) involve an equal number of experimental units. In other words, each factor level will have the same number of exposed units. When the set of levels is exhaustive and not merely a sample, such a factor's effects are *fixed*.

With a within-subjects factor, each experimental unit is exposed to all factor levels. Such a factor's effect is *random* when it involves a random sampling of the set of levels that can be found for some larger population. Blocking effects may be treated as random when the experimental blocks are seen as a sample of the population of blocks about which inferences are to be drawn.

Consider this example: An EXPERIMENT tests, by gender, the effect of drug therapy on 100 severely depressed patients. The gender factor in the study has only two levels, male and female, both of which have equal numbers. Gender here is a between-subjects factor with fixed effects. The drug therapy factor, on the other hand, is based on a random selection of four drugs (levels) from the large set of antidepressant drugs now available for prescription. Such sampling permits estimation of the random (treatment) effect of the entire population of antidepressant drugs in dealing with depression. *Drug* here is a within-subjects factor. Note, however, that had the design involved not testing an overall (sampled) drug effect but rather the individual effect for each of the four drugs—and *only* these four—the drug factor effects would be considered fixed, not random. In either case, when the selection of subjects for the study can reasonably be assumed to be a random selection from a larger POPULATION of patients, *patient* also would be a factor with random effects.

There are numerous types of mixed designs. Although many are quite similar, each may require somewhat different data-analytic methods, depending on how treatments are applied to experimental units. Four common mixed designs are REPEATED MEASURES, SPLIT PLOT, NESTED or hierarchical, and RANDOMIZED BLOCKS. In *repeated-measures designs* (sometimes called within-subjects designs), experimental units are measured more than once. When a time series of effect data will be collected, measurement intervals are to be of equal length. These designs are fully randomized and may or may not include a between-subjects factor (other than the treatment). If such a factor is present, the design is said to have a split-plot/repeated-measures structure.

In basic *split-plot designs*, experimental units are of two different sizes, the larger being blocks of the smaller. Two treatments are involved. The first (the within-subjects factor) is assigned to the larger units, and the second (the between-subjects factor) is assigned to the smaller units separately within each

larger block. There are similarities between repeated-measures designs and split-plot designs in that the treatment factor in the former is equivalent to the larger unit ("whole plot") treatment in the latter. In addition, the time factor in the repeated-measures design is equivalent to the smaller unit ("subplot") treatment in the split-plot design. However, the covariance structure of the data from a repeated-measures design usually requires methods of analysis different from those for data from a true split-plot design.

Nested or hierarchical designs involve the nesting of one experimental factor within another. (Nested designs are sometimes called between-subjects designs.) A simple version of this design is found when experimental units merely are nested within conditions. In a "purely hierarchical" design, every set of factors has an inside-outside relationship. When blocks are nested within a between-blocks factor, the nested design is equivalent to a split-plot design. With nesting of blocks, the inside factor groups are completely embedded within the larger factor grouping. Consequently, it is not possible to separate the insider factor's effects from an insider-outsider interaction.

In *randomized blocks designs*, treatments are randomly assigned to units within blocks. Treatment effects are usually considered fixed (no treatments other than those actually in the experiment are considered). Block effects are considered random, a sampling of the larger set of blocks about which conclusions are to be drawn. In cases when each treatment is applied in each block, the design is called a randomized complete blocks design.

The analysis of the data from mixed designs is made problematic by the random factor(s). The recently developed and now preferred approach uses MAXIMUM LIKELIHOOD ESTIMATION (rather than LEAST SQUARES), in which there is no assumption that variance components are independently and identically distributed. This is possible because the covariances are modeled within the analysis itself. Such mixed modeling can generate appropriate error terms even in the presence of MISSING DATA (and an unbalanced design). Nonetheless, valid SIGNIFICANCE TESTS for the random factors still must meet the SPHERICITY ASSUMPTION for *F* tests to be valid. Sphericity (or its equivalents: circularity, Huynh-Feldt structure, or Type H covariance structure) is found when all the variances of the differences (in the population sampled) are equal.

—Douglas Madsen

REFERENCES

- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute, Inc.
- Milliken, G. A., & Johnson, D. E. (1992). *Analysis of messy data: Vol. 1. Designed experiments*. London: Chapman & Hall.

MIXED-EFFECTS MODEL

ANALYSIS OF VARIANCE (ANOVA) models can be used to study how a dependent variable Y depends on one or several factors. A *factor* here is defined as a categorical independent variable—in other words, an explanatory variable with a nominal LEVEL OF MEASUREMENT. Factors can have fixed or random effects. For a factor with fixed effects, the parameters expressing the effect of the categories of each factor are fixed (i.e., nonrandom) numbers. For a factor with random effects, the effects of its categories are modeled as random variables. Fixed effects are appropriate when the statistical inference aims at finding conclusions that hold for precisely the categories of the factor that are present in the data set at hand. Random effects are appropriate when the categories of the factor are regarded as a random sample from some population, and the statistical inference aims at finding conclusions that hold for this population. If the model contains only fixed or only random effects, we speak of FIXED-EFFECTS MODELS or RANDOM-EFFECTS MODELS, respectively. A model containing fixed as well as random effects is called a mixed-effects model.

An example of a mixed-effects ANOVA model could be a study about work satisfaction of employees in several job categories in several organizations. If the researcher is interested in a limited number of specific job categories and in a population of organizations, as well as, inter alia, the amount of variability between the organizations in this population, then it is natural to use a mixed model, with fixed effects for the job categories and random effects for the organizations.

Random effects, as defined above and in the entry on random-effects models, can be regarded as random main effects of the levels of a given factor. Random interaction effects can be defined similarly. In the example from the previous paragraph, a random job

category-by-organization interaction effect also could be included.

Mixed-effects models are also used for the analysis of data collected according to SPLIT-PLOT DESIGNS or REPEATED-MEASURES DESIGNS. In such designs, individual subjects (or other units) are investigated under several conditions. Suppose that the conditions form one factor, called B ; the individual subjects are also regarded as a factor, called S . The mixed-effects model for this design has a fixed effect for B , a random effect for S , and a random $B \times S$ interaction effect.

THE LINEAR MIXED MODEL

A generalization is a random coefficient of a numerical explanatory variable. This leads to the generalization of the GENERAL LINEAR MODEL, which incorporates fixed as well as random coefficients, commonly called the linear mixed model or random coefficient regression model. In matrix form, this can be written as

$$Y = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \mathbf{E},$$

where Y is the dependent variable, $\mathbf{X}$ is the design matrix for the fixed effects, $\boldsymbol{\beta}$ is the vector of fixed regression coefficients, $\mathbf{Z}$ is the design matrix for the random coefficients, $\mathbf{U}$ is the vector of random coefficients, and $\mathbf{E}$ is the vector of random residuals. The standard assumption is that the vectors $\mathbf{U}$ and $\mathbf{E}$ are independent and have multivariate normal distributions with zero mean vectors and with covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively. This assumption implies that Y has the multivariate normal distribution with mean $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$. Depending on the study design, a further structure will usually be imposed on the matrices $\mathbf{G}$ and $\mathbf{R}$. For example, if there are K factors with random effects and the residuals are purely random, then the model can be specialized to

$$Y = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{U}_1 + \mathbf{Z}_2\mathbf{U}_2 + \cdots + \mathbf{Z}_K\mathbf{U}_K + \mathbf{E},$$

where now all vectors $\mathbf{U}_1, \dots, \mathbf{U}_K$ and $\mathbf{E}$ are independent and have multivariate normal distributions with zero mean vectors and covariance matrices $\mathbf{G}_1, \dots, \mathbf{G}_K$ and $\sigma^2\mathbf{I}$, respectively, and where no further restrictions are imposed on the matrices $\mathbf{G}_1$ to $\mathbf{G}_K$ and the scalar σ^2 . This model is the matrix formulation of the HIERARCHICAL LINEAR MODEL but without the assumption that the factors with random effects are hierarchically nested, and it is a basic model for MULTI-LEVEL ANALYSIS. For $K = 1$, this is the two-level

hierarchical linear model, often used for the analysis of longitudinal data (see Verbeke & Molenberghs, 2000, and the entry on MULTILEVEL ANALYSIS). For longitudinal data, the covariance matrix $\mathbf{R}$ can also be assumed to have specific nondiagonal forms, expressing autocorrelation or other kinds of time dependence.

Mixed-effects models can also be defined as extensions of the GENERALIZED LINEAR MODEL. In such models, the linear predictor $\eta = \mathbf{X}\beta$, as defined in the entry on generalized linear models, is replaced by the term $\mathbf{X}\beta + \mathbf{Z}\mathbf{U}$. This gives a general mathematical formulation of so-called generalized linear mixed models, which generalize HIERARCHICAL NONLINEAR MODELS to the case of random factors that are not necessarily nested.

—Tom A. B. Snijders

REFERENCES

- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Longford, N. T. (1993). *Random coefficient models*. New York: Oxford University Press.
- McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear and mixed models*. New York: John Wiley.
- Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. W. (1996). *Applied linear regression models*. New York: Irwin.
- Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

MIXED-METHOD RESEARCH. See MULTIMETHOD RESEARCH

MIXTURE MODEL (FINITE MIXTURE MODEL)

Finite mixture model is another name for LATENT CLASS ANALYSIS. These models assume that parameters of a statistical model of interest differ across unobserved subgroups called *latent classes* or *mixture components*. The most important kinds of applications of finite mixture models are clustering, scaling, density estimation, and RANDOM-EFFECTS MODELING.

—Jeroen K. Vermunt

MLE. See MAXIMUM LIKELIHOOD ESTIMATION

MOBILITY TABLE

A mobility table is a two-way CONTINGENCY TABLE in which the individuals are classified according to the same variable on two occasions. It therefore distinguishes between the “immobile” (i.e., all the cases on the main diagonal of the table) and the “mobile,” who are classified elsewhere. Mobility tables are very common in the social sciences (e.g., to study geographical mobility of households from one census to another or to examine electorate mobility between candidates during an election campaign). In sociology, intergenerational mobility tables cross-classify the respondents’ own social class with their class of origin (usually father’s class), whereas intragenerational mobility tables cross-classify their occupation or class at two points in time.

HISTORICAL DEVELOPMENT

Essentially developed in sociology from the 1950s, the statistical methodology of mobility tables was primarily based on the computation of indexes. The mobility ratio or Glass’s index was proposed as the ratio of each observed frequency in the mobility table to the corresponding expected frequency under the hypothesis of statistical independence (also called “perfect mobility”). Yasuda’s index and Boudon’s index were based on the decomposition of total mobility as the sum of two components: structural or “forced” mobility (the dissimilarity between marginal distributions implies that all cases cannot be on the diagonal) and net mobility (conceived of as reflecting the intensity of association). However, with time, the shortcomings of mobility indexes became increasingly clear.

Since the 1980s, the new paradigm has been to analyze mobility tables from two perspectives. Total immobility rate ($\sum_i f_{ii}/f_{++}$), outflow percentages (f_{ij}/f_{i+}), and inflow percentages (f_{ij}/f_{+j}) are simple tools to describe absolute (or observed) mobility. But absolute mobility also results from the combination of the marginals of the table with an intrinsic pattern

of ASSOCIATION between the variables. The study of relative mobility consists of analyzing the structure and strength of that association with LOG-LINEAR MODELS and ODDS RATIOS.

Among these models, “topological” or “levels” models are an important class based on the idea that different densities underlie the cells of the mobility table. In the two-way multiplicative (log-linear) model $F_{ij} = \alpha\beta_i\gamma_j\delta_{ij}$, positing $\delta_{ij} = 1$ for all i and all j leads to the independence model. It assumes that a single density applies for the entire table; that is, each of the mobility or immobility trajectories is equally plausible (net of marginal effects). Positing $\delta_{ij} = \delta$ if $i = j$ and $\delta_{ij} = 1$ otherwise corresponds to a two-density model (with one degree of freedom less than the independence model) called the *uniform inheritance model* because δ , which is usually more than 1, captures a propensity toward immobility common to all categories. Assuming that this propensity is specific to each category ($\delta_{ij} = \delta_i$ if $i = j$) leads to the *quasi-perfect mobility model*. With the insight that the same density can apply to cells on and off the diagonal, R. M. Hauser (1978), following previous work by Leo Goodman,

proposed the general levels model: $F_{ij} = \alpha\beta_i\gamma_j\delta_k$ if the cell (i, j) belongs to the k th density level.

EXAMPLE

Table 1 is a mobility table that cross-classifies the father’s occupation by the son’s first full-time civilian occupation for U.S. men between ages 20 and 64 in 1973. The total immobility rate is 39.9% $((1,414 + 524 + \dots + 1,832)/19,912)$; that is, two out of five sons belong to the same category as their father. As regards the outflow perspective, 39.4% of farmers’ sons are themselves farmers $(1,832/4,650)$; as regards the inflow perspective, 80.9% of farmers have farm origins $(1,832/2,265)$. These two figures and their contrast reflect (and confound) both the interaction effect (the propensity toward immobility into farming) and marginal effects (the importance of the farm category among fathers and among sons).

To isolate interaction effects, we need log-linear modeling. Although the perfect mobility model is very far from the data (the LIKELIHOOD RATIO STATISTIC L^2 is 6,170.13 with 16 degrees of freedom), the

Table 1 Mobility Table

Father's Occupation	Son's Occupation					Total
	Upper Nonmanual	Lower Nonmanual	Upper Manual	Lower Manual	Farm	
Upper nonmanual	1,414	521	302	643	40	2,920
Lower nonmanual	724	524	254	703	48	2,253
Upper manual	798	648	856	1,676	108	4,086
Lower manual	756	914	771	3,325	237	6,003
Farm	409	357	441	1,611	1,832	4,650
Total	4,101	2,964	2,624	7,958	2,265	19,912

Table 2 Five-Level Model of Mobility

Father's Occupation	Son's Occupation				
	Upper Nonmanual	Lower Nonmanual	Upper Manual	Lower Manual	Farm
Upper nonmanual	B	D	E	E	E
Lower nonmanual	C	D	E	E	E
Upper manual	E	E	E	E	E
Lower manual	E	E	E	D	D
Farm	E	E	E	D	A

uniform inheritance model yields a huge improvement ($L^2 = 2,567.66$, $df = 15$, $\delta = 2.62$). So, there is a marked inheritance effect. However, its strength varies from one category to another because the quasi-perfect mobility model is still much better ($L^2 = 683.34$, $df = 11$). Finally, Table 2 presents a levels model proposed by Featherman and Hauser (1978), which reproduces the observed data fairly well ($L^2 = 66.57$, $df = 12$). Given that $\delta_E = 1$, the estimated interaction parameters are $\delta_A = 30.04$, $\delta_B = 4.90$, $\delta_C = 2.47$, and $\delta_D = 1.82$. Thus, for instance, propensity toward immobility is the highest in the farm category, and the upward move from lower nonmanual to upper nonmanual is more probable than the corresponding downward move (net of marginal effects).

ADVANCED TOPICS

Although the levels model can be applied when the variables are nominal or ordinal, special models are also useful in the latter case: diagonals models, crossings parameters models, uniform association model, and row (and/or column) effect models. Log-multiplicative ASSOCIATION MODELS are relevant when the categories are assumed to be ordered, but their exact order is unknown and must be established from the data.

When several mobility tables are grouped together according to a third (layer) variable (e.g., country or time), log-linear modeling of the variation in the association over that dimension usually generates many supplementary parameters. Recent progress has been made in modeling such a variation very parsimoniously with log-multiplicative terms. These methods have, for instance, shown that the general strength of association between class of origin and destination has steadily decreased in France for men and women from the 1950s to the 1990s.

—Louis-André Vallet

REFERENCES

- Featherman, D. L., & Hauser, R. M. (1978). *Opportunity and change*. New York: Academic Press.
- Hauser, R. M. (1978). A structural model of the mobility table. *Social Forces*, 56, 919–953.
- Hout, M. (1983). *Mobility tables*. Beverly Hills, CA: Sage.
- Vallet, L.-A. (2001). Forty years of social mobility in France. *Revue Française de Sociologie: An Annual English Selection*, 42(Suppl.), 5–64.

- Xie, Y. (1992). The log-multiplicative layer effect model for comparing mobility tables. *American Sociological Review*, 57, 380–395.

MODE

The mode is a MEASURE OF CENTRAL TENDENCY and refers to the value or values that appear most frequently in a distribution of values. Thus, in the series 9, 8, 2, 4, 9, 5, 2, 7, 10, there are two modes (2 and 9).

—Alan Bryman

MODEL

Social scientists use *model* to denote two distinct (although related) concepts: *empirical models* and *theoretical models*. An empirical model usually takes the form of an empirical test, often a statistical test, such as in a REGRESSION that is used to discern whether a set of independent variables influences a DEPENDENT VARIABLE. In such empirical models, the INDEPENDENT VARIABLES are identified from either a theory or observed empirical regularities.

A theoretical model, like an empirical model, represents an abstraction from reality. The goal, however, is different. In an empirical model, the analyst seeks to ascertain which independent variables actually exert a significant influence on the dependent variable of interest. The empirical test then indicates which of the potential influences are actual influences and which are not. In a theoretical model, on the other hand, the analyst seeks to provide an abstract account of how various forces interact to produce certain outcomes. The goal of such models is to capture what is sometimes referred to as the data-generating process, which is the source for data that exist in the real world (Morton, 2000). To capture this process, the analyst will create an argument, or “tell a story,” about how real-world outcomes or events occur and, in so doing, will identify the key factors and mechanisms that will lead to these outcomes or events.

Some theoretical models are purely verbal and use words to construct their arguments. Others, however, use mathematics to construct their arguments. Generally, models that rely on mathematics to provide predictions about the world are known as *formal*

models. These models consist of several components. First, there are constants, or undefined terms, which are those features of the model that are taken as given (e.g., members of a legislature). Second, there are defined terms (e.g., committees consist of members of the legislature). Third, there are assumptions about goals and preferences. For example, an analyst might assume that committees are driven by policy concerns. Finally, there is a sequence, which details the order in which events in the model take place (e.g., a bill is introduced in a legislature, a committee decides whether to send the bill to the chamber floor, etc.).

Once these various pieces are laid out, the analyst attempts to solve the model—that is, to identify the conditions under which different outcomes occur. Models can be solved in a variety of ways. GAME-THEORETIC models, for example, are usually solved analytically, in which the analyst identifies a solution, or equilibrium. Dynamic models, on the other hand, are usually solved computationally, in which the analyst sets initial values for the variables and then runs SIMULATIONS to provide insights about outcomes.

Formal models offer several benefits over verbal, or nonformal, models. First, the assumptions in formal models need to be stated more clearly. Second, the terms are defined more specifically. Third, the logic of the argument is spelled out more cleanly and transparently.

—Charles R. Shipan

REFERENCE

Morton, Rebecca B. (2000). *Methods and models: A guide to the empirical analysis of formal models in political science*. New York: Cambridge University Press.

MODEL I ANOVA

ANALYSIS OF VARIANCE (ANOVA) is a method that deals with differences in means of a variable across categories of observations. A classic example would be the comparison of a behavior of interest between an EXPERIMENTAL group and a control group. The question is whether the mean behavior (measured on an INTERVAL scale) of the subjects in the experimental group differs significantly from the mean behavior of the members of the control group. In other words, does the experimental treatment have an effect? ANOVA is a method that employs ratios of variances—hence

the name *analysis of variance*—to conduct a test of SIGNIFICANCE. So long as we have observations from all categories, as opposed to a subset of categories, ANOVA uses a FIXED-EFFECTS MODEL. In the example at hand, there are only two categories—namely, the experimental condition and the control condition. So, the fixed-effects model would apply. Note, however, that the fixed-effects model is not restricted to comparisons across just two categories.

ANOVA results typically appear in a table with the following layout (Table 1):

Table 1

Source	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio	Significance
Between groups	10.0	1	10.0	1.21	.30
Within groups	66.0	8	8.3		
Total	76.0	9			

SOURCE: Iversen and Norpoth (1987, p. 21).

The “sum of squares” entries provide measures of variation between the groups and within each of the groups, with the assumption that the within-group variation is constant. To standardize those measures, we divide each sum of squares by the respective degrees of freedom. That gives us two estimates of VARIANCE (MEAN SQUARE). In the event that the between variance is no larger than the within variance, the ratio of those two quantities (*F* RATIO) should be 1.0. In that case, any difference between the two groups is entirely due to random chance and not to any group characteristic. In the case above, the *F* ratio only reaches 1.21, which is not significant at the .05 level. Hence, we have to conclude that the experimental treatment has no effect.

The fixed-effects model is not limited to testing for the effect of just one explanatory variable (ONE-WAY ANOVA). It can also accommodate two such variables (TWO-WAY ANOVA), in which one is designated as the row variable and the other as the column variable. It is advisable, however, to design the research in such a way that all the resulting combinations (cells) of rows and columns contain the same number of observations. That is possible, of course, only with experiments in which the investigators have that kind of control. With survey or other observational data, combinations of categories will have whatever observations occur. In that case, one cannot obtain unique estimates for the variance due to each of the explanatory factors. That

limits the utility of two-way ANOVA to experimental designs.

Besides testing for the effect of each explanatory variable (MAIN EFFECTS), a two-way ANOVA also provides a test of the INTERACTION between those variables. Two variables (X_1 and X_2) are said to have an interaction effect on the dependent variable (Y) if the main effects are not additive. In other words, the addition of those effects does not predict the dependent variable of interest, even aside from random error. Being exposed to two experimental conditions has an effect on a behavior of interest that is systematically higher (or lower) than what the two separate effects add up to. With categorical data, such interaction effects are often difficult to interpret in the absence of good theory about their directions.

—Helmut Norpoth

See also FIXED-EFFECTS MODEL

REFERENCES

- Hays, W. L. (1994). *Statistics* (5th ed.). New York: Harcourt Brace.
- Iversen, G. R., & Norpoth, H. (1987). *Analysis of variance* (2nd ed., Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-001). Newbury Park, CA: Sage.
- Scheffe, H. (1959). *The analysis of variance*. New York: John Wiley.

MODEL II ANOVA. See
RANDOM-EFFECTS MODEL

MODEL III ANOVA. See
MIXED-EFFECTS MODEL

MODELING

Modeling is the process of creating “a simplified representation of a system or phenomenon . . . with any hypotheses required to describe the system or explain the phenomenon” (Random House, 1997). Through

the modeling process, analysts seek to confirm or deny the HYPOTHESES they generate about the system or phenomenon of concern, usually accomplishing the task of confirmation or denial through the testing of the hypotheses using QUANTITATIVE DATA. The analyst hopes to arrive at conclusions from the exercise that are nontrivial, scientifically sound, and replicable using the same MODEL and data. The conclusions should also help both the analyst and the reader to explain what took place, predict future behavior, and present fresh conceptualizations that can better explain and predict societal behavior.

HISTORY

Model building became most prevalent in social science research during the behaviorist revolution that took place in social science beginning in the mid-20th century. At that time, social science scholars, inspired by the mathematical and logical sophistication of the physical sciences (biology, chemistry, physics), especially of economics, became particularly concerned with finding “universal” relationships between quantitatively measured variables that were similar to those found by physical scientists. The purpose behind seeking universal relationships was to build lawlike statements about human behavior. The variance in human behavior, particularly in the social science realm, was to be accounted for through the use of scientifically applied control variables that operated independently of the human behavior under study. Blalock (1969) considered inductive model building to be essential for more precise theorizing and, if necessary, for making adjustments to the original proposed theory based on the results of hypothesis testing. Over the years, studies using more deductive formal models that depend on the logical sequencing of arguments have become more popular and have taken their place alongside their inductive cousins in the social science literature.

As scientists in the physical sciences know the precise amount of acceleration to apply to a mass to produce a known quantity of force, expressed in the formula $f = m \times a$, behaviorist-minded social scientists could also know to a more refined extent the necessary amount of an INDEPENDENT VARIABLE to create an important impact on the DEPENDENT VARIABLE, which is the human behavior of interest. Thus, *voting behavior* became a term of choice in political science and political sociology because social

scientists increasingly understood the power of models, within which several independent variables could be placed, to capture variation in individual voting choices at election time. Each independent variable in the model accounts for a different portion of the variance.

MODEL BUILDING

Social scientists build models to better understand the “real-world” environment. In particular, issues of cause and effect are addressed through modeling. What explains the number of votes cast for Candidate Brown for mayor? How can we account for changes in regime type in countries around the world? What factors are most important in explaining variation in spending on social welfare programs in the U.S. states?

Social science modeling creates an artificial and abstract societal environment that represents its most important features. For instance, social scientists attempt to account for the vote choice by ascertaining the most critical factors that voters consider in helping them to choose to vote for a particular candidate or against another candidate or candidates: party identification, ideology, and opinions about the condition of the economy. Then, through the measurement of these important (though hardly exhaustive list of) variables, the analyst can focus his or her attention on comprehending what is essential about the real-world environment. This should bring about a better understanding of the phenomenon under study (i.e., account for variation in vote choice).

TYPES OF MODELS

Models can be both conceptual and mathematical. The goal of the social scientist is to proceed from the conceptual model to the mathematical model. The difference between the two is that although conceptual models allow the analyst to play with competing theories in verbal form, mathematical models are made “OPERATIONAL” using variables that numerically measure a phenomenon of interest. The mathematical measures can be either direct or indirect. An example of a direct measure is votes cast for a particular candidate, which represents the support a candidate has garnered to win elective office. An example of an indirect measure is an index of democracy, in which different components of the measure are assembled to represent the concept of “democracy.” Models further specify a path by which the variables influence one another to

produce a behavior. Thus, mathematical models are “testable,” in that one can estimate them using statistical techniques and come to conclusions about the hypotheses about the real world that inform them.

CONSTRUCTING AND EVALUATING MODELS

For instance, to better understand the occurrence of democratic regimes in different countries, a researcher could create a conceptual model in the following manner:

$$\text{Democracy} = f(\text{Economics, Culture}).$$

The researcher could further refine his or her effort by conceiving of variables that operationalize the above concepts as follows:

$$\begin{aligned} \text{Democracy} = f(\text{Economic Development,} \\ \text{Income Distribution, Religious} \\ \text{Practice, Ethnic Composition}). \end{aligned}$$

In the model, economic development and income distribution (throughout the population of the country) represent “Economics,” whereas religious practice (e.g., the percentage of the population that is Protestant) and ethnic composition of the country’s population represent “Culture.”

The modeler also hypothesizes, in this model, that the relationship between the independent variables and the dependent variable is linear and direct. Models can also be constructed that specify relationships that are linear and indirect (such as INTERACTION terms), nonlinear and direct (through a LOGARITHMIC TRANSFORMATION), and NONLINEAR and indirect.

How do we evaluate models? Do they accurately explain variance in the dependent variable of interest? If we estimate models using statistics, we generally use statistical standards to evaluate their performance. The multivariate nature of modeling lends itself to several standards. The *R-SQUARED* statistics is a common evaluation statistic for MULTIPLE REGRESSION models. The log-likelihood is another way to judge multivariate model performance—in this instance, MAXIMUM LIKELIHOOD ESTIMATION.

Models are created to answer specific research questions that are more comprehensive in scope than can be captured by a simple bivariate relationship. Models help social scientists to explain, predict, and

understand societal phenomena. They can also be evaluated for their effectiveness in performing those three tasks.

—Ross E. Burkhart

REFERENCES

- Asher, H. B. (1983). *Causal modeling* (2nd ed., Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-003). Beverly Hills, CA: Sage.
- Blalock, H. M., Jr. (1969). *Theory construction: From verbal to mathematical formulations*. Englewood Cliffs, NJ: Prentice Hall.
- Lave, C., & March, J. G. (1978). *An introduction to models in the social sciences*. New York: Harper & Row.
- Random House. (1997). *Random House Webster's college dictionary*. New York: Author.
- Schrodt, P. A. (2002). Mathematical modeling. In J. B. Manheim, R. C. Rich, & L. Willnat (Eds.), *Empirical political analysis: Research methods in political science* (5th ed.). New York: Longman.

MODERATING. See MODERATING VARIABLE

MODERATING VARIABLE

A moderating variable represents a process or a factor that alters the impact of an independent variable X on a dependent variable Y . The moderating variable may take the form of an indicator variable (0/1 values), a categorical variable, or a continuous variable. Moderation of the effects of X on Y is usually of interest in the context of a well-theorized area of research, such as the provision of labor by individuals in which the human capital and labor market segmentation theories are both well established. Variables reflecting these two theories can be put together in a multiple regression equation, and a path analysis involving two stages (labor force participation in the first stage and wage determination in the second) can be developed. Paths through the model involve the multiplication of the coefficients from two regressions together, giving the moderated effect of X on Y through Z , and comparing that effect with the direct effect of X on Z (which is assumed to be additively separable) (Bryman & Cramer, 1997).

Furthermore, background factors can be tested for a moderating effect on the main impact of each X on the final outcome Y .

Moderating variables allow for testing the notion of contingent effects. Realists distinguish necessary effects from contingent effects (Sayer, 2000, p. 94). Researchers benefit from focusing on contingent effects that only occur in the absence of offsetting, blocking, or cross-cutting causal processes. The moderating variables may act as obstacles to the action of X in causing Y . Thus, divorce only results from "submitting divorce papers" *if* the case is well constructed, uncontested, or both; it is contingently caused by divorce papers.

Two main approaches are used for estimating the impact of moderating variables: PATH ANALYSIS and INTERACTION EFFECTS.

The path analysis approach allows for two or more equations to simultaneously represent the set of associations. These equations may involve independent variables X_{ij} , intermediate variables Z_{ij} , and final outcomes Y_{ij} . In each equation j , variables may be recursively affected by factors in other linked equations. Such models are structural equation models.

The alternative method of using interaction effects (Jaccard, Turrisi, & Wan, 1990) introduces an extra term that is the product of two existing variables into a single regression equation. The interaction terms may be numerous for categorical variables. The degrees of freedom of the new estimate are reduced.

For example, in researching the factors associated with eating out frequently, the interaction effect of marital status with having young children in the household was measured using dummy variables in a logistic regression context (Olsen, Warde, & Martens, 2000). The impact of [Divorced $\times$ Kids < 5 in Hhold] was a dramatic increase in the odds of eating out in a hotel or of eating an unusual ethnic meal. This interaction term offset one of the underlying associations: Divorced people in the United Kingdom were generally less likely than married/cohabiting people to eat in ethnic restaurants.

A simpler alternative is to use dummy variables to control for the impact of moderating factors. Another alternative is to use multilevel modeling to represent contextual moderating influences (Jones, 1997).

—Wendy K. Olsen

REFERENCES

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bryman, A., & Cramer, D. (1997). *Quantitative data analysis with SPSS for Windows: A guide for social scientists*. New York: Routledge.
- Jaccard, J., Turrissi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression* (Quantitative Applications in the Social Sciences, No. 72). London: Sage.
- Jones, K. (1997). Multilevel approaches to modelling contextuality: From nuisance to substance in the analysis of voting behaviour. In G. P. Westert & R. N. Verhoeff (Eds.), *Places and people: Multilevel modelling in geographical research* (Nederlandse Geografische Studies 227). Utrecht, The Netherlands: The Royal Dutch Geographical Society and Faculty of Geographical Sciences, Utrecht University.
- Olsen, W. K., Warde, A., & Martens, L. (2000). Social differentiation and the market for eating out in the UK. *International Journal of Hospitality Management*, 19(2), 173–190.
- Sayer, A. (2000). *Realism and social science*. London: Sage.

MOMENT

Moment is often used with an order value in statistical analysis. The k th moment of a variable is the average value of the observations of the variable raised to the k th power. Let N indicate sample size and the cases of the variable be indicated by $x_1, x_2, \dots, x_N$, and then the k th moment of the variable is

$$(x_1^k + x_2^k + x_N^k)/N.$$

The k th moment of a random variable X is the expected value of X^k , or $E(X^k)$.

In qualitative research, the word *moment* often is used to refer to a specific combination of social and historical circumstances such as a sociohistorical moment.

—Tim Futing Liao

MOMENTS IN QUALITATIVE RESEARCH

The history of qualitative research in North America in the 20th century divides into seven phases, or moments. We define them as the *traditional* (1900–1950); the *modernist* or golden age (1950–1970); *blurred genres* (1970–1986); the *crisis of representation* (1986–1990); the *postmodern*, a period

of experimental and new ethnographies (1990–1995); *postexperimental* inquiry (1995–2000); and the *future*, which is now. The future, the seventh moment, is concerned with moral discourse; with the rejoining of art, literature, and science; and with the development of sacred textualities. The seventh moment asks that the social sciences and the humanities become sites for critical conversations about democracy, race, gender, class, nation, globalization, freedom, and community.

THE TRADITIONAL PERIOD

The first moment, the traditional period, begins in the early 1900s and continues until World War II. During this time, qualitative researchers wrote “objective,” colonizing accounts of field experiences, reflective of the POSITIVIST scientist paradigm. They were concerned with offering VALID, RELIABLE, and OBJECTIVE interpretations in their writings. The other who was studied was often alien, foreign, and exotic.

MODERNIST PHASE

The modernist phase, or second moment, builds on the canonical works from the traditional period. Social REALISM, NATURALISM, and slice-of-life ETHNOGRAPHIES are still valued. It extended through the postwar years to the 1970s and is still present in the work of many. In this period, many texts sought to formalize and legitimize qualitative methods. The modernist ethnographer and sociological PARTICIPANT OBSERVER attempted rigorous, qualitative studies of important social processes, including deviance and social control in the classroom and society. This was a moment of creative ferment.

BLURRED GENRES

By the beginning of the third stage (1970–1986), “blurred genres,” qualitative researchers had a full complement of paradigms, methods, and strategies to employ in their research. Theories ranged from SYMBOLIC INTERACTIONISM to CONSTRUCTIVISM, naturalistic inquiry, positivism and postpositivism, PHENOMENOLOGY, ETHNOMETHODOLOGY, critical Marxist theories, SEMIOTICS, STRUCTURALISM, FEMINISM, and various ethnic paradigms. APPLIED QUALITATIVE RESEARCH was gaining in stature, and the politics and ethics of qualitative research were topics of considerable concern. Research strategies and formats for reporting research ranged from GROUNDED

THEORY to CASE STUDY to methods of historical, biographical, ethnographic, action, and clinical research. Diverse ways of collecting and analyzing empirical materials became available, including qualitative interviewing (open-ended and semistructured), observational, visual, personal experience, and documentary methods. Computers were entering the situation, to be fully developed (to provide online methods) in the next decade, along with narrative, content, and semiotic methods of reading interviews and cultural texts. Geertz's two books, *The Interpretation of Cultures* (1973), and *Local Knowledge* (1983), defined the beginning and end of this moment.

CRISIS OF REPRESENTATION

A profound rupture occurred in the mid-1980s. What we call the fourth moment, or the crisis of representation, appeared with *Anthropology as Cultural Critique* (Marcus & Fischer, 1986), *The Anthropology of Experience* (Turner & Bruner, 1986), and *Writing Culture* (Clifford & Marcus, 1986). They articulated the consequences of Geertz's "blurred genres" interpretation of the field in the early 1980s as new models of truth, method, and representation were sought (Rosaldo, 1989).

THE POSTMODERN MOMENT

The fifth moment is the postmodern period of experimental ethnographic writing. Ethnographers struggle to make sense of the previous disruptions and crises. New ways of composing ethnography are explored. Writers struggle with new ways to represent the "other" in the text. Epistemologies from previously silenced groups have emerged to offer solutions to these problems. The concept of the aloof observer has been abandoned.

THE POSTEXPERIMENTAL AND BEYOND MOMENTS

The sixth (postexperimental) and seventh (future) moments are upon us. Fictional ethnographies, ethnographic poetry, and multimedia texts are today taken for granted. Postexperimental writers seek to connect their writings to the needs of a free democratic society. The needs of a moral and sacred qualitative social science are actively being explored by a host of new writers (Clough, 1998).

We draw several conclusions from this brief history, noting that it is, like all histories, somewhat arbitrary. First, North Americans are not the only scholars struggling to create postcolonial, nonessentialist, feminist, dialogic, performance texts—texts informed by the rhetorical, narrative turn in the human disciplines. This international work troubles the traditional distinctions between science, the humanities, rhetoric, literature, facts, and fictions. As Atkinson and Hammersley (1994), observe, this discourse recognizes "the literary antecedents of the ethnographic text, and affirms the essential dialectic" (p. 255) underlying these aesthetic and humanistic moves.

Second, this literature is reflexively situated in multiple, historical, and national contexts. It is clear that America's history with qualitative inquiry cannot be generalized to the rest of the world (Atkinson, Coffey, & Delamont, 2001). Nor do all researchers embrace a politicized, cultural studies agenda that demands that interpretive texts advance issues surrounding social justice and racial equality.

Third, each of the earlier historical moments is still operating in the present, either as legacy or as a set of practices that researchers continue to follow or argue against. The multiple and fractured histories of qualitative research now make it possible for any given researcher to attach a project to a canonical text from any of the above-described historical moments. Multiple criteria of evaluation compete for attention in this field. Fourth, an embarrassment of choices now characterizes the field of qualitative research. There have never been so many paradigms, strategies of inquiry, or methods of analysis to draw on and use. Fifth, we are in a moment of discovery and rediscovery, as new ways of looking, interpreting, arguing, and writing are debated and discussed. Sixth, the qualitative research act can no longer be viewed from within a neutral or objective positivist perspective. Class, race, gender, and ethnicity inevitably shape the process of inquiry, making research a multicultural process.

—Norman K. Denzin and Yvonna S. Lincoln

REFERENCES

- Atkinson, P., Coffey, A., & Delamont, S. (2001). Editorial: A debate about our canon. *Qualitative Research*, 1, 5–21.
- Atkinson, P., & Hammersley, M. (1994). Ethnography and participant observation. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 248–261). Thousand Oaks, CA: Sage.

- Clifford, J., & Marcus, G. E. (Eds.). (1986). *Writing culture*. Berkeley: University of California Press.
- Clough, P. T. (1998). *The end(s) of ethnography* (2nd ed.). New York: Peter Lang.
- Geertz, C. (1973). *The interpretation of cultures*. New York: Basic Books.
- Geertz, C. (1983). *Local knowledge*. New York: Basic Books.
- Marcus, G., & Fischer, M. (1986). *Anthropology as cultural critique*. Chicago: University of Chicago Press.
- Rosaldo, R. (1989). *Culture & truth*. Boston: Beacon.
- Turner, V., & Bruner, E. (Eds.). (1986). *The anthropology of experience*. Urbana: University of Illinois Press.

MONOTONIC

Monotonic refers to the properties of a function that links one set of scores to another set of scores. Stated formally, consider a set of scores Y that are some function of another set of scores, X , such that $Y = f(X)$. The function is said to be *monotonic increasing* if for any two scores, X_j and X_k , where $X_j < X_k$, the property holds that $f(X_j) \leq f(X_k)$. If $f(X_j) < f(X_k)$, then the function is said to be *monotonic strictly increasing*. If $f(X_j) \geq f(X_k)$, then the function is said to be *monotonic decreasing*, and if $f(X_j) > f(X_k)$, then the function is said to be *monotonic strictly decreasing*.

For example, consider five individuals on whom a measure of an X variable is taken as well as five additional variables, A , B , C , D , and E . The data are as follows in Table 1.

Table 1

Individual	X	A	B	C	D	E
1	12	44	44	47	47	47
2	13	45	44	46	46	46
3	14	46	45	45	45	45
4	15	47	46	44	44	46
5	16	48	47	43	44	47

A is a monotonic strictly increasing function of X because for any two individuals where X_j is less than X_k , it is the case that $A_j < A_k$. B is a monotonic increasing function of X because for any two individuals where X_j is less than X_k , it is the case that $B_j \leq B_k$. C is a monotonic strictly decreasing function of X because for any two individuals where X_j is less than X_k , it is the case that $C_j > C_k$. D is a monotonic decreasing function of X because for any two individuals where X_j is less than X_k , it is the case

that $B_j \geq B_k$. Finally, E is a nonmonotonic function of X because it satisfies none of the conditions described above.

—James J. Jaccard

MONTE CARLO SIMULATION

Monte Carlo simulation is the use of computer ALGORITHMS and pseudo-random numbers to conduct mathematical experiments on statistics. These experiments are used to understand the behavior of statistics under specified conditions and, ultimately, to make inferences to population characteristics. Monte Carlo simulation is most useful when strong analytic distributional theory does not exist for a statistic in a given situation.

EXPLANATION

The goal of INFERENCE STATISTICAL analysis is to make PROBABILITY statements about a POPULATION PARAMETER, θ , from a statistic, $\hat{\theta}$, calculated from a SAMPLE of data from a POPULATION. This is problematic because the calculated statistics from any two RANDOM SAMPLES from the same population will almost never be equal. For example, suppose you want to know the average IQ of the 20,000 students at your university (θ), but you do not have the time or resources to test them all. Instead, you draw a random sample of 200 and calculate the mean IQ score ($\hat{\theta}$) of these sampled students to be, say, 123. What does that tell you about the average IQ of the entire student body? This is not obvious, because if you drew another random sample of 200 from those same 20,000 students, their mean IQ might be 117. A third random sample might produce a mean of 132, and so on, through the very large number of potential random samples that could be drawn from this student body. So how can you move from having solid information about the sample statistic ($\hat{\theta}$) to being able to make at least a probability statement about the population parameter (θ)?

At the heart of making this inference from sample to population is the SAMPLING DISTRIBUTION. A statistic's sampling distribution is the range of values it could take on in a random sample from a given population and the probabilities associated with those values. So the mean IQ of a sample of 200 of our 20,000 students might be as low as, say, 89 or as high as 203, but there would

probably be a much greater likelihood that the mean IQ of such a sample would be between 120 and 130. If we had some information about the structure of the sample mean's sampling distribution, we might be able to make a probability statement about the population mean, given a particular sample.

In the standard parametric inferential statistics that social scientists learn in graduate school (with the ubiquitous t -tests and p values), we get information about a statistic's sampling distribution from mathematical analysis. For example, the CENTRAL LIMIT THEOREM gives us good reason to believe that the sampling distribution of the mean IQ of our random sample of 200 students is distributed normally, with an expected value of the population mean and a STANDARD DEVIATION of approximately the standard deviation of the IQ in the population divided by the square root of 200. However, there are situations when either no such analytical distributional theory exists about a statistic or the assumptions needed to apply parametric statistical theory do not hold. In these cases, one can use Monte Carlo simulation to estimate the sampling distribution of a statistic empirically.

Monte Carlo simulation estimates a statistic's sampling distribution empirically by drawing a large number of random samples from a known "pseudo-population" of simulated data to track a statistic's behavior. The basic concept is straightforward: If a statistic's sampling distribution is the density function of the values it could take on in a given population, then the relative frequency distribution of the values of that statistic that are actually observed in many samples from a population is an estimate of that sampling distribution. Because it is impractical for social scientists to sample actual data multiple times, we use artificially generated data that resemble the real thing in relevant ways for these samples.

The basic steps of a Monte Carlo simulation are as follows:

1. Define in symbolic terms a pseudo-population that will be used to generate random samples. This usually takes the form of a computer algorithm that generates simulated data in a carefully specified manner.
2. Sample from the pseudo-population (yielding a pseudo-sample) in ways that simulate the statistical situation of interest, for example, with the same sampling scheme, sample size, and so on.
3. Calculate and save $\hat{\theta}$ from the pseudo-sample.
4. Repeat Steps 2 and 3 t times, where t is the number of trials. Usually, t is a large number, often 10,000 or more. This is done with a looping algorithm in a computer program.
5. Construct a relative frequency distribution of the resulting $\hat{\theta}_t$ values. This is the Monte Carlo estimate of the sampling distribution of $\hat{\theta}$ under the conditions specified by the pseudo-population and the sampling procedures.
6. Evaluate or use this estimated sampling distribution to understand the behavior of $\hat{\theta}$ given a population value, θ .

Thus, Monte Carlo simulation is conducted by writing and running a computer program that generates carefully structured, artificial, random data to conduct a mathematical experiment on the behavior of a statistic. Two crucial steps in the process are (a) specifying the pseudo-population in a computer algorithm and (b) using the estimated sampling distribution to understand social processes better.

Specifying the pseudo-population with a computer algorithm must begin with a clear understanding of the social process to be simulated. Such a computer algorithm will be as complex as the process being simulated is believed to be. Most theoretical models of social processes contain a deterministic and a probabilistic, or random, component. Writing the computer code for the deterministic component is often relatively straightforward. The random component of the simulation usually requires the most thought. This is appropriate because the random component of a social process largely determines the configuration of the statistics' sampling distributions and makes statistical inference an issue in the first place. It is the importance of the random component in this technique that led to it being named for the famous European casino.

The random component enters a Monte Carlo simulation of a social process in the distributions of the variables that go into it. In writing a simulation program, you specify these distributions based on theory and evidence about the phenomena you are modeling and the characteristics of the experiment you are undertaking. In theory, a social variable could take on any of an infinite range of distributions, but statisticians have identified and developed theory about relatively few. Most books on Monte Carlo simulation, including those listed below, discuss at length how to

generate variables with various distributions and what observed variables might be best simulated by each of them.

APPLICATION

Monte Carlo simulation is based on a tradition of numerical experimentation that goes back at least to biblical times, but with renewed interest being sparked in it in the mid-20th century from such disparate fields as quantum physics and the development of telephone networks. However, it was not until very recently that the technique has become widely practical, due to the massive computer power available for the first time at the end of the 20th century. As a result, social scientists are only just beginning to develop general uses for Monte Carlo simulation (but the rise in its use has been rapid, with a key word search for *Monte Carlo* in the Social Sciences Citation Index yielding only 14 articles in 1987 compared to 270 articles in 2001). Some of the most important current applications of Monte Carlo simulation include the following:

- assessing the sampling distribution of a new statistic (or combination of statistics) for which mathematical analysis is intractable;
- assessing the sensitivity of an empirical analysis to variations in potential population conditions, including the violation of parametric inference assumptions; and
- being used in MARKOV CHAIN MONTE CARLO techniques, which have recently brought Bayesian statistical inference into the social science mainstream.

See Evans, Hastings, and Peacock (2000); Mooney (1997); Robert and Casella (2000); and Rubinstein (1981) for these and other uses of Monte Carlo simulation in the social sciences.

—Christopher Z. Mooney

REFERENCES

- Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical distributions* (3rd ed.). New York: John Wiley.
- Mooney, C. Z. (1997). *Monte Carlo simulation*. Thousand Oaks, CA: Sage.
- Robert, C. P., & Casella, G. (2000). *Monte Carlo statistical methods*. New York: Springer-Verlag.
- Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo method*. New York: John Wiley.

MOSAIC DISPLAY

Mosaic displays (or mosaic plots) are plots of cell frequencies in a cross-classified CONTINGENCY TABLE depicting the cell frequency $n_{ijkl\dots}$ as the area of a “tile” in a space-filling mosaic graphic. This is based on the idea that, for a two-way table, cell probabilities $p_{ij} = n_{ij}/n_{++}$ can always be expressed as $p_{ij} = p_{i+} \times p_{j|i}$, shown by the width and height of rectangles.

Thus, if a unit area is divided first in proportion to the marginal probabilities of one variable, p_{i+} , and those are each subdivided according to conditional probabilities, $p_{j|i}$, the tiles (a) will have area $\sim p_{ij}$ and (b) will align under independence (i.e., when $p_{ij} = p_{i+} \times p_{+j}$). This visual representation turns out to have a long and interesting history (Friendly, 2002).

Recently reintroduced to statistical graphics (Hartigan & Kleiner, 1981), the basic mosaic display generalizes this idea to n -way tables by recursive subdivision of areas according to the continued application of Bayes rule, so that $p_{ijkl\dots} = p_i \times p_{j|i} \times p_{k|ij} \times \dots \times p_{n|ijk\dots}$.

This method for visualizing contingency tables was enhanced (Friendly, 1992, 1994, 1998) to (a) fit any log-linear model, (b) show the residuals (contributions to the overall Pearson or likelihood ratio chi-square) from the model by color and shading, and (c) permute the categories of variables so that the pattern of shading shows the nature of the associations among variables. As a consequence, mosaic displays have become a primary graphical tool for visualization and analysis of categorical data in the form of n -way contingency tables (Friendly, 1999, 2000; Valero, Young, & Friendly, 2003).

A two-way table and a three-way table example in Figure 1 show the relations among the categories of hair color, eye color, and sex. In these figures, the area of each tile is proportional to the observed cell frequency, but the tiles are shaded according to the residuals from a particular log-linear model, thus showing the pattern of associations that remain. (Color versions are far more effective and use shades of blue and red for positive and negative residuals, respectively.) In the left-hand two-way display, the model of independence has been fit to the table; the opposite-corner pattern of the shading shows that dark (light) hair is associated with dark (light) eyes.

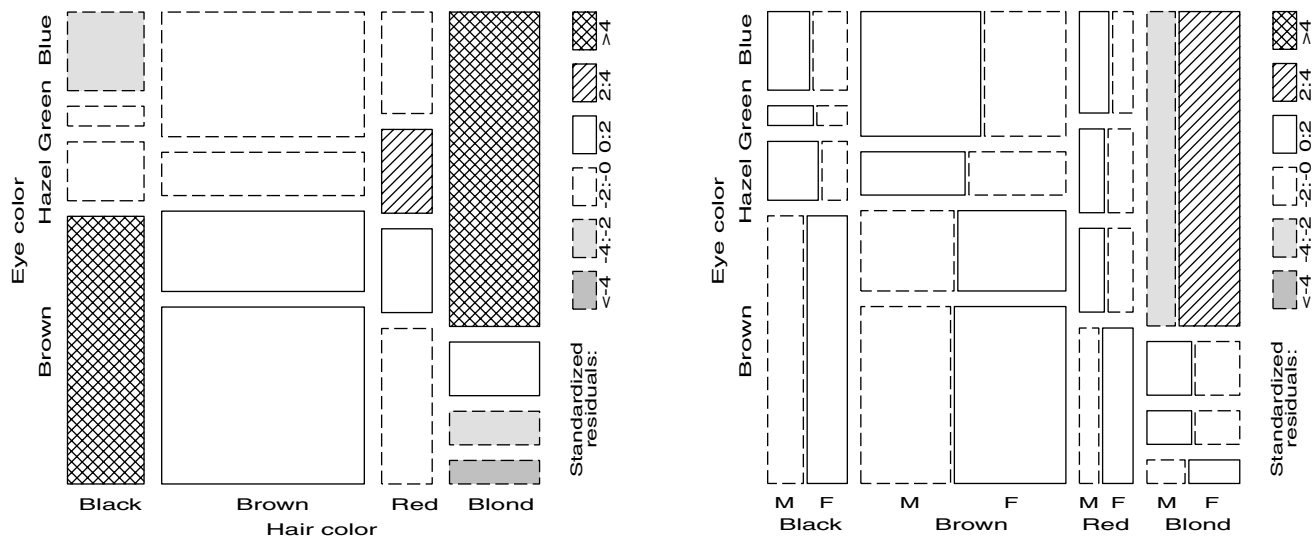


Figure 1 Mosaic Displays for Frequencies of Hair Color, Eye Color, and Sex

NOTE: Left: two-way table, hair color by eye color; right: three-way table, divided by sex.

In the right panel, the log-linear model [*Hair Eye*] [*Sex*] has been fit to the three-way table, which asserts that the combinations of hair color and eye color are independent of sex. Only one pair of cells has residuals large enough to be shaded—the blue-eyed blondes—in which there are more women and fewer men than would be the case under this model of independence.

Recent developments extend mosaic displays in several directions: (a) mosaic matrices and partial mosaics (Friendly, 1999, 2000) as discrete analogs of scatterplot matrices and coplots for continuous data; (b) several highly interactive, dynamic implementations (Hofmann, 2000; Theus, 1997; Valero et al., 2003) for exploration and “visual fitting”; and (c) application of same visual ideas to nested classifications or tree structures (Shneiderman, 1992).

—Michael Friendly

REFERENCES

- Friendly, M. (1992). Mosaic displays for loglinear models. In American Statistical Association (Ed.), *Proceedings of the Statistical Graphics Section* (pp. 61–68). Alexandria, VA: American Statistical Association.
- Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89, 190–200.
- Friendly, M. (1998). Mosaic displays. In S. Kotz, C. Reed, & D. L. Banks (Eds.), *Encyclopedia of statistical sciences* (Vol. 2, pp. 411–416). New York: John Wiley.

- Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3), 373–395.
- Friendly, M. (2000). *Visualizing categorical data*. Cary, NC: SAS Institute.
- Friendly, M. (2002). A brief history of the mosaic display. *Journal of Computational and Graphical Statistics*, 11(1), 89–107.
- Hartigan, J. A., & Kleiner, B. (1981). Mosaics for contingency tables. In W. F. Eddy (Ed.), *Computer science and statistics: Proceedings of the 13th Symposium on the Interface* (pp. 268–273). New York: Springer-Verlag.
- Hofmann, H. (2000). Exploring categorical data: Interactive mosaic plots. *Metrika*, 51(1), 11–26.
- Shneiderman, B. (1992). Tree visualization with treemaps: A 2-D space-filling approach. *ACM Transactions on Graphics*, 11(1), 92–99.
- Theus, M. (1997). Visualization of categorical data. In W. Bandilla & F. Faulbaum (Eds.), *SoftStat'97: Advances in statistical software* (Vol. 6, pp. 47–55). Stuttgart, Germany: Lucius & Lucius.
- Valero, P., Young, F., & Friendly, M. (2003). Visual categorical analysis in ViSta. *Computational Statistics and Data Analysis*, 43(4), 495–508.

MOVER-STAYER MODELS

The mover-stayer (MS) model is an extension of the MARKOV CHAIN MODEL for dealing with a very specific type of unobserved heterogeneity in the population.

The population is assumed to consist of two unobserved groups (latent classes): a stayer group containing persons with a zero probability of change and a mover group following an ordinary Markov process. Early references to the MS model are Blumen, Kogan, and McCarthy (1955) and Goodman (1961). Depending on the application field, the model might be known under a different name. In the biomedical field, one may encounter the term *long-term survivors*, which refers to a group that has a zero probability of dying or experiencing another type of event of interest within the study period. In marketing research, one uses the term *brand-loyal segments* as opposed to *brand-switching segments*. There is also a strong connection with zero-inflated models for binomial or Poisson counts.

Suppose the categorical variable Y_t is measured at T occasions, where t denotes a particular occasion, $1 \leq t \leq T$; D is the number of levels of Y_t ; and y_t is a particular level of Y_t , $1 \leq y_t \leq D$. This could, for example, be a respondent's party preference measured at 6-month intervals. The vector notation $\mathbf{Y}$ and $\mathbf{y}$ is used to refer to a complete response pattern.

A good way to introduce the MS model is as a special case of the mixed Markov model. The latter is obtained by adding a discrete unobserved heterogeneity component to a standard Markov chain model. Let X denote a discrete latent variable with C levels. A particular latent class is enumerated by the index x , $x = 1, 2, \dots, C$. The mixed Markov model has the form

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^C P(X = x) P(Y_1 = y_1 | X = x) \times \prod_{t=2}^T P(Y_t = y_t | Y_{t-1} = y_{t-1}, X = x).$$

As can be seen, latent classes differ with respect to the initial state and the transition probabilities. A MS model is obtained if $C = 2$ and $P(Y_t = y_t | Y_{t-1} = y_t, X = 2) = 1$ for each y_t . In other words, the probability of staying in the same state equals 1 in latent class 2, the stayer class. Note that $P(X = x)$ is the proportion of persons in the mover and stayer classes, and $P(Y_1 = y_1 | X = x)$ is the initial distribution of these groups over the various states.

To identify the MS model, we need at least three occasions, but it is not necessary to assume the process to be stationary in the mover chain. With three time points ($T = 3$) and two states ($D = 2$), a nonstationary MS model is a saturated model.

A restricted variant of the MS model is the independence-stayer model, in which the movers are assumed to behave randomly; that is, $P(Y_t = y_t | Y_{t-1} = y_{t-1}, X = 1) = P(Y_t = y_t | X = 1)$. Another variant is obtained with absorbing states, such as first marriage and death. In such cases, everyone starts in the first state, $P(Y_1 = 1 | X = x) = 1$, and backward transitions are also impossible in the mover class, $P(Y_t = 2 | Y_{t-1} = 2, X = 1) = 1$.

The formulation as a mixed Markov model suggests several types of extensions, such as the inclusion of more than one class of movers or a LATENT MARKOV MODEL variant that corrects for measurement error. Another possible extension is the inclusion of covariates, yielding a model that is similar to a zero-inflated model for counts.

Two programs that can be used to estimate the MS model and several of its extensions are PANMARK and LEM.

—Jeroen K. Vermunt

REFERENCES

- Blumen, I. M., Kogan, M., & McCarthy, P. J. (1955). *The industrial mobility of labor as a probability process*. Ithaca, NY: Cornell University Press.
- Goodman, L. A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56, 841–868.
- Langeheine, R., & Van de Pol, F. (1994). Discrete-time mixed Markov latent class models. In A. Dale & R. B. Davies (Eds.), *Analyzing social and political change: A casebook of methods* (pp. 171–197). London: Sage.
- Vermunt, J. K. (1997). *Log-linear models for event histories*. Thousand Oaks, CA: Sage.

MOVING AVERAGE

A moving average process is the product of a RANDOM shock, or disturbance, to a TIME SERIES. When the pattern of disturbances in the dependent variable can best be described by a weighted average of current and lagged random disturbances, then it is a moving average process. There are both immediate effects and discounted effects of the random shock over time, occurring over a fixed number of periods before disappearing. In a sense, a moving average is a linear combination of WHITE-NOISE error terms. The number of periods that it takes for the effects of the random shock on the dependent variable to disappear is termed

the ORDER of the moving average process, denoted q . The length of the effects of the random shock is not dependent on the time at which the shock occurs, making a moving average process independent of time.

The mathematical representation of a moving average process with mean $\mu = 0$ and of order q is as follows:

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}.$$

Take, for instance, a moving average process of order $q = 1$: The effects of the random shock in period t would be evident in period t and period $t - 1$, then in no period after. Although observations one time period apart are correlated, observations more than one time period apart are uncorrelated. The memory of the process is just one period. The order of the process can be said to be the memory of the process. An example would be the effect of a news report about the price of hog feed in Iowa. The effect of this random shock would be felt for the first day and then discounted for 1 or more days (order q). After several days, the effects of the shock would disappear, resulting in the series returning to its mean.

Identifying a moving average process requires examining the AUTOCORRELATION (ACF) and partial autocorrelation (PACF) functions. For a first-order moving average process, the ACF will have one spike and then abruptly cut off, with the remaining lags of the function exhibiting the properties of white noise; the PACF of a first-order moving average process will dampen infinitely. Figure 1 shows the ACF for a first-order moving average process. Moving average processes of order q will have an ACF that spikes significantly for lags 1 through q and then abruptly cuts off after lag q . In contrast, the PACF will dampen infinitely. Statistical packages such as SAS and SPSS will generate correlograms for quick assessment of ACFs and PACFs, although higher-order moving average processes are rare in the social sciences.

A moving average process can occur simultaneously with serial correlation and a TREND. Autoregressive, integrated, moving average (ARIMA) modeling is useful to model time series that exhibit a combination of these processes because it combines an autoregressive process, an integrated or differenced process (which removes a trend), and a moving average process. Furthermore, a moving average process can be written as an infinite-order autoregressive process. Conversely, an infinite-order autoregressive process can be written as a first-order moving average process. This is

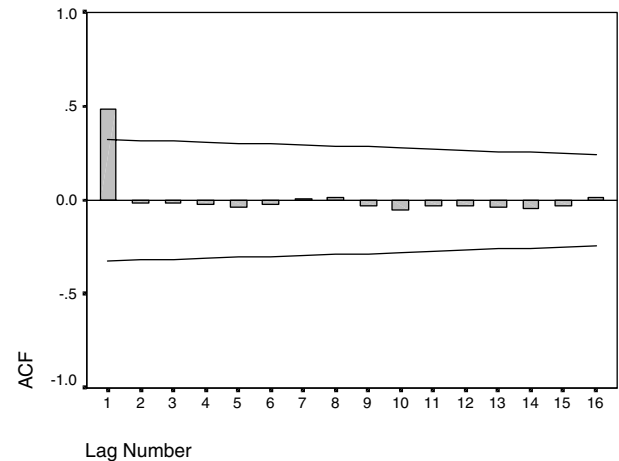


Figure 1 Theoretical Autocorrelation (ACF): MA(1) Process

important because it allows for parsimonious expressions of complex processes as small orders of the other process.

—Andrew J. Civettini

REFERENCES

- Cromwell, J. B., Labys, W., & Terraza, M. (1994). *Univariate tests for time series models*. Thousand Oaks, CA: Sage.
 Ostrom, C. W., Jr. (1990). *Time series analysis*. Thousand Oaks, CA: Sage.

MPLUS

Mplus is a statistical software package for estimating STRUCTURAL EQUATION MODELS and performing analyses that integrate RANDOM-EFFECTS FACTOR, and LATENT CLASS ANALYSIS in both cross-sectional and longitudinal settings and for both single-level and multilevel data. For further information, see the following Web site of the software: www.statmodel.com/index2.html.

—Tim Futing Liao

MULTICOLLINEARITY

Collinearity between two INDEPENDENT VARIABLES or multicollinearity between multiple independent variables in LINEAR REGRESSION analysis means that there are linear relations between these variables. In

a vector model, in which variables are represented as vectors, exact collinearity would mean that two vectors lie on the same line or, more generally, for k variables, that the vectors lie in a subspace of a dimension less than k (i.e., that one of the vectors is a linear combination of the others). The result of such an exact linear relation between variables $\mathbf{X}$ would be that the cross-product matrix $\mathbf{X}'\mathbf{X}$ is singular (i.e., its determinant would be equal to zero), so that it would have no inverse. In practice, an exact linear relationship rarely occurs, but the interdependence of social phenomena may result in "approximate" linear relationships. This phenomenon is known as multicollinearity.

In the analysis of interdependence, such as PRINCIPAL COMPONENTS and FACTOR ANALYSIS, multicollinearity is blissful because dimension reduction is the objective, whereas in the analysis of dependence structures, such as MULTIPLE REGRESSION ANALYSIS, multicollinearity among the independent variables is a problem. We will treat it here in the context of regression analysis.

CONSEQUENCES

Imagine an investigation into the causes of well-being of unemployed workers (measured as a scale from 0 to 100). A great number of EXPLANATORY VARIABLES will be taken into account, such as annual income (the lower the annual income, the lower the well-being), duration of unemployment (the higher the number of weeks of unemployment, the lower the well-being), and others. Now, some of these independent variables will be so highly correlated that conclusions from a multiple regression analysis will be affected. High multicollinearity is undesirable in a multicausal model because standard errors become much larger and CONFIDENCE INTERVALS much wider, thus affecting SIGNIFICANCE TESTING and STATISTICAL INFERENCE, even though the ORDINARY LEAST SQUARES (OLS) estimator still is unbiased. Like VARIANCES and COVARIANCES, multicollinearity is a sample property. This means analyses may be unreliable and sensitive to a few OUTLIERS or extreme cases.

The problem of correlated independent variables in a causal model is an old sore in social scientific research. In multiple regression analysis, it is referred to as multicollinearity (the predictors in a vector model almost overlap in a straight line), whereas in ANALYSIS OF VARIANCE and COVARIANCE, one tends to use the term *nonorthogonality* (the factors of the

factor-response model are not perpendicular to each other). *Multicollinearity* and *nonorthogonality* are, in fact, synonyms.

Let us emphasize that strange things may happen in an analysis with high multicollinearity. Kendall and Stuart (1948) provided an extreme example of one dependent variable Y and two independent variables X_1 and X_2 , in which the correlations $Y - X_1$ and $Y - X_2$ are weak (one of them even zero) and the correlation $X_1 - X_2$ is very strong (0.90, high multicollinearity), whereas the multiple correlation coefficient $R_{y.12}$ is approximately 1. This classic example shows a classic consequence of multicollinearity: insignificant results of testing individual parameters but high R -SQUARED values. The standardized coefficients (a direct function of sample variances), which normally cannot be greater than 1 because their squared values express a proportion of explained variance, exceed 2; consequently, both variables appear to explain each more than 400%, with only 100% to be explained!

DETECTIONS

How do we know whether the analysis is affected by high multicollinearity?

Most statistical software packages such as SPSS, SAS, and BMDP contain "collinearity diagnostics." Correlations with other independent variables, multiple correlations from the auxiliary regression of X_j on all other independent variables (R_j), TOLERANCE ($1 - R_j^2$), variance-inflation factors [$1/(1 - R_j^2)$], and eigenvalue analyses (cf. Belsley, Kuh, & Welsch, 1980) are part of the standard output and indicate which subset of variables is problematic. These diagnoses are more appropriate than the suggestion to draw the line at an arbitrary correlation value such as 0.60, as is stated in some textbooks. Besides, these diagnoses are to be used together with a SIGNIFICANCE TEST of the regression coefficients (or a calculation of the confidence intervals). Also, one should not take the correlation among X variables as proof enough for multicollinearity, for one should not forget that a correlation has to be evaluated in a multivariate context; that is, it might be affected by spurious relations, suppressor variables, and the like.

SOLUTIONS

A first option, ignoring the problem, can only be justified if multicollinearity only refers to subsets of

variables, the individual effects of which the researcher is not interested in.

An obvious solution would be to delete one of these highly correlated variables. However, researchers often do not find this approach satisfactory from a theoretical point of view. To take the aforementioned example, both annual income and duration of unemployment offer a particular and fairly distinct explanation for well-being, even if they are statistically correlated.

A third possible solution, as suggested by econometricians, is to get more data. By obtaining more data, the current sample and its properties are modified. However, this solution is impractical most of the time because extra time and money are needed.

A fourth procedure removes the common variance of one collinear variable from another. In the current example, this would imply that the researcher keeps annual income but only keeps the residual of the duration of unemployment after regressing it on income. This procedure can be applied to all variables involved.

A fifth method is the creation of a summary index of the collinear variables if the researcher is not interested in seeing separate effects of the variables involved. But this method ignores measurement error and other possible factors.

A more sophisticated approach is the use of principal components analysis (PCA). All independent variables (often a great number) are to be used as the basis for a PCA and possibly combined with VARIMAX rotation, from which the component scores are taken and included in further regression analysis with, say, well-being as the dependent variable and the components as independent variables. The independent variables then will be at perpendicular angles to each other, thus avoiding multicollinearity. Interpretation of the new components can be a challenge, although the problem can be avoided by conducting the PCA on collinear variables only.

Finally, when the researcher intends to keep the highly correlated variables X_1 and X_2 in the regression model and yet produce a scientifically well-founded analysis, we can recommend *Ridge Regression* (e.g., Hoerl & Kennard, 1970) and *Bayes Regression* (e.g., Lindley & Smith, 1972), although they do not provide a 100% satisfactory solution either.

Ridge regression is an attempt to cope with multicollinearity by introducing some bias in the OLS estimation. Estimates from ridge regression have lower standard errors, but they are also slightly biased. In a $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ regression, the coefficients $\hat{\boldsymbol{\beta}}^* =$

$(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$ are used, in which small increments to the principal diagonal ($\mathbf{X}'\mathbf{X}$) are added. Now the entire system behaves more orthogonally. However, the choice of the small values k is arbitrary. Higher k values decrease standard errors but introduce greater bias in the estimates. Ridge regression is available in major statistical software programs today.

Similar to ridge regression, Bayes regression attempts to correct the problems of multicollinearity. However, while ridge regression relies on an arbitrary choice by the researcher of the small increments k , these are to be estimated from the data in Bayes regression. The disadvantage is that the Bayes strategy invariably presupposes the formulation of an a priori distribution.

—Jacques Tacq

REFERENCES

- Belsley, D. A., Kuh, E., & Welsch, R. (1980). *Regression diagnostics*. New York: John Wiley.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression. *Technometrics*, 12(1), 55–67, 69–82.
- Kendall, M. G., & Stuart, A. (1948). *The advanced theory of statistics*. London: Griffin.
- Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society*, 34, 1–41.

MULTIDIMENSIONAL SCALING (MDS)

Multidimensional scaling (MDS) refers to a family of models in which the structure in a set of data is represented graphically by the relationships between a set of points in a space. MDS can be used on a wide variety of data, using different models and allowing different assumptions about the level of measurement.

In the simplest case, a data matrix giving information about the similarity (or DISSIMILARITY) between a set of objects is represented by the proximity (or distance) between corresponding points in a low-dimensional space. Given a set of data, interpreted as “distances,” it finds the map locations that generated them.

For example, in a study of how people categorize drugs, a list of drugs was elicited from a sample of users and nonusers; 28 drugs were retained for a free-SORTING experiment, and the co-occurrence frequency was used as the measure of similarity. The

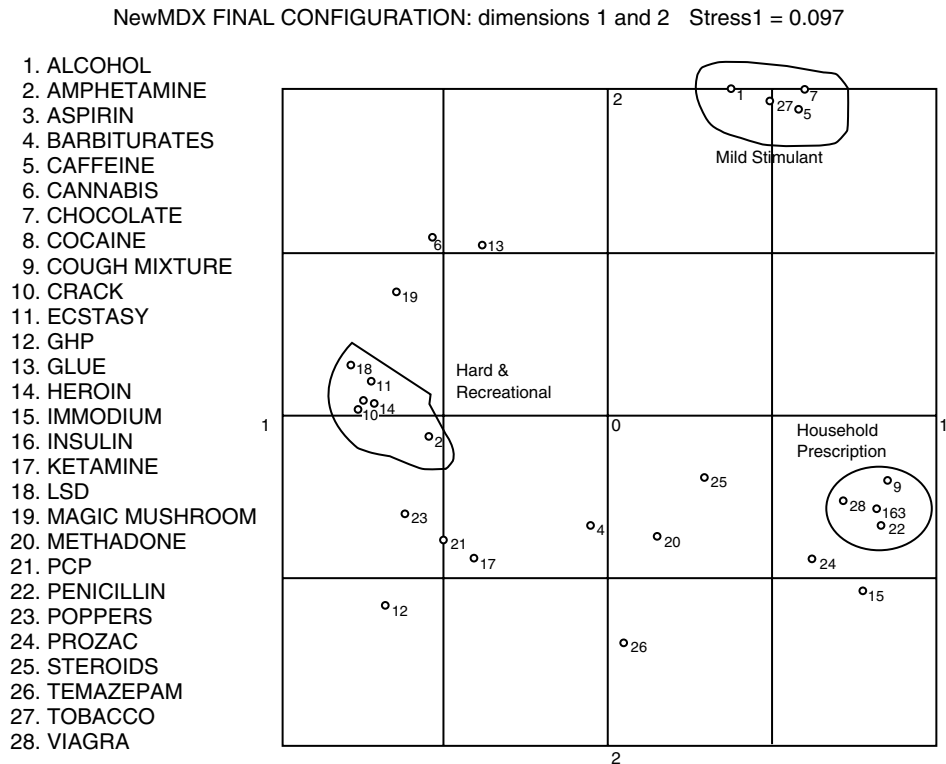


Figure 1

data were scaled in two dimensions, producing Figure 1.

In the case of perfect data, the correspondence between the dissimilarity data and the distances of the solution will be total, but with imperfect data, the degree of fit is given by the size of the normalized stress (residual sum of squares) value, which is a measure of badness of fit. In this case, the fit is excellent ($\text{stress}_1 = 0.097$), and it is three times smaller than random expectation. The solution configuration yields three highly distinct clusters (confirmed by independent hierarchical clustering) of *mild stimulant drugs* (core prototypes: tobacco, caffeine), *hard recreational* (core: cocaine, heroin), and *household prescription* (core: aspirin, penicillin).

MDS can be either exploratory (simply providing a useful and easily assimilable visualization of a data set) or explanatory (giving a geometric representation of the structure in a data set, in which the assumptions of the model are taken to represent the way in which the data were produced). Compared to other multivariate methods, MDS models are usually distribution free, make conservative demands on the structure of the data, are unaffected by nonsystematic missing data, and can

be used with a wide variety of measures; the solutions are also usually readily interpretable. The chief weaknesses are relative ignorance of the sampling properties of stress, proneness to local minima solutions, and inability to represent the asymmetry of causal models.

THE BASIC MDS MODEL

The basic type of MDS is the analysis of two-way, one-mode data (e.g., a matrix of correlations or other dissimilarity/similarity measures), using the Euclidean distance model. The original *metric* version, “classical” MDS converted the “distances” into scalar products and then factored or decomposed them into a set of locations in a low-dimensional space (hence “smallest space analysis”). The first *nonmetric* model was developed by Roger N. Shepard in 1962, who showed that the merely ordinal constraints of the data, if imposed in sufficient number, guarantee metric recovery, and he provided the first iterative computer program that implemented the claim. Kruskal (1964) gave ordinary least squares (OLS) statistical substance to it; Takane, Young, and DeLeeuw (1977) developed a frequently used Alternating Least Squares

program (ALSCAL); and probabilistic MAXIMUM LIKELIHOOD versions of MDS have also subsequently been developed (e.g., MULTISCALE) (Ramsay, 1977).

VARIANTS OF MDS

The various types of MDS can be differentiated (Coxon, 1982) in terms of the following:

- **The data**, whose “shape” is described in terms of its *way* (rows, columns, “third-way” replications) and its *mode* (the number of sets of distinct objects, such as variables, subjects).
- **The transformation (rescaling) function**, or LEVEL OF MEASUREMENT, which specifies the extent to which the data properties are to be adequately represented in the solution. The primary distinction is *nonmetric* (ordinal or lower) versus *metric* (interval and ratio) scaling.
- **The model**, which is usually the Euclidean distance model but also covers simpler Minkowski metrics such as City-Block and the different composition functions such as the vector, scalar products, or FACTOR model.

TWO-WAY, ONE-MODE DATA

Any nonnegative, symmetric judgments or measures can provide input to the basic model: frequencies, counts, ratings/rankings of similarity, co-occurrence, colocation, confusion rates, measures of association, and so forth, and a two-dimensional solution should be stable with 12 or more points. Nonmetric analyses use distance models; metric models also may use vector/factor models. Stress values from simulation studies of random configurations provide an empirical yardstick for assessing obtained stress values.

TWO-WAY, TWO-MODE DATA

Rectangular data matrices (usually with subjects as rows and variables as columns) consist of profile data or preference ratings/rankings. Distance (unfolding) or vector models are employed to produce a joint BILOT of the row and column elements as points. The vector model (MDPREF) is formally identical to simple CORRESPONDENCE ANALYSIS and represents the row elements as unit vectors. Because such data are row conditional, caution is needed in interpreting inter-set relations in the solution. If the sorting data of the

example were entered directly as (Individual $\times$ Object matrix, with entries giving the category number), then a program such as MDSORT would represent both the objects and the categories.

THREE-WAY DATA

The “third way” consists of different individuals, time or spatial points, subgroups in the data, experimental conditions, different methods, and so forth. The most common variant is three-way, two-mode data (a stack of two-way, one-mode matrices)—which, when represented by the metric weighted distance model, is termed *individual differences scaling* (INDSCAL) (Carroll & Chang, 1970). In the INDSCAL solution, the objects are represented in a group space (whose dimensions are fixed and should be readily interpretable), and each element of the third way (“individual”) has a set of nonnegative dimensional weights by which the individual differentially shrinks or expands each dimension of the group space to form a “private space.” Individual differences are thus represented by a profile of salience weights applied to a common space. When plotted separately as the “subject space,” the angular separation between the points represents the closeness of their profiles, and the distance from the origin approximately represents the variance explained. Returning to the example, if the individuals were divided into subgroups (such as Gender $\times$ Usage) and a co-occurrence matrix calculated within each group, then the set of matrices forms three-way, two-mode data, and their scaling by INDSCAL would yield information about how far the groups differed in terms of the importance (weights) attributed to the dimensions underlying the drugs configuration.

Other variants of MDS exist for other types of data (e.g., tables, triads), other transformations (parametric mapping, power functions), and other models (non-Euclidean metrics, simple conjoint models, discrete models such as additive trees), as well as for a wide range of three-way models and hybrid models such as CLASCAL (a development of INDSCAL that parameterizes latent *classes* of individuals). Utilities exist for comparing configuration (Procrustean rotation), and extensions for up to seven-way data are possible.

An extensive review of variants of MDS is contained in Carroll and Arabie (1980), and a wide-ranging bibliography of three-way applications is accessible at <http://ruls01.fsw.leidenuniv.nl/~kroonenb/document/reflist.htm>.

APPLICATIONS

For the basic model and for the input of aggregate measures, restrictions on the number of cases are normally irrelevant. But for two-way, two-mode data and three-way data, applications are rarely feasible (or interpretable) for more than 100 individual cases. In this case, "pseudo-subjects" (subgroups of individuals chosen either on analytic grounds or after CLUSTER ANALYSIS) are more appropriately used. Disciplines using MDS now extend well beyond the original focus in psychology and political science and include other social sciences, such as economics, as well as biological sciences and chemistry.

—Anthony P. M. Coxon

REFERENCES

- Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via a N-way generalisation of Eckart-Young decomposition. *Psychometrika*, 35, 283–299, 310–319.
- Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology*, 31, 607–649.
- Coxon, A. P. M. (1982). *The user's guide to multidimensional scaling*. London: Heinemann Educational.
- Kruskal, J. B. (1964). Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–27, 115–129.
- Ramsay, J. O. (1977). Maximum likelihood estimation in MDS. *Psychometrika*, 42, 241–266.
- Takane, Y., Young, F. W., & DeLeeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42, 7–67.

MULTIDIMENSIONALITY

A unidimensional structure exists in those situations in which a single, fundamental DIMENSION underlies a set of observations. Examples of this condition are measures used to evaluate the qualifications of political candidates, a SCALE of political activism, or a measure designed to describe a single personality trait. The presumption is that there is a single common dimension on which the candidates, citizens, or personality traits are arrayed.

Many theoretical CONSTRUCTS used in the social sciences are quite complex. A single, homogeneous, underlying dimension may not be adequate to express these complex concepts. This complexity may require us to turn to more elaborate explanations of the

variables' behavior in which more than one latent dimension underlies a set of empirical observations. This is multidimensionality. For these concepts to be useful in further analysis, their multiple aspects must be reflected in their MEASUREMENT. In this regard, specifying multiple dimensions is a way of defining the complex concept by simplifying it into measurable and understandable aspects and may not represent substantive reality. Its utility is in OPERATIONALIZING the theoretical construct for ease of interpretation.

The substantive task of the researcher relative to dimensionality is to uncover these multiple dimensions and determine how many are scientifically important. We hypothesize their existence because THEORY suggests that they should exist or that we can predict these latent dimensions because of the variation in the empirical observations at hand. We hypothesize how they should relate to each other and then use some empirical technique to test our HYPOTHESES. Thus, there is a theoretical and an empirical basis to the concept.

For example, modeling the concept of political ideology by using a variable measuring only attitudes toward governmental intervention in the economy gives an entirely different picture of ideology than if we defined it by measuring only attitudes toward issues of race, such as busing or affirmative action. We could further model ideology by measuring attitudes on social-cultural issues such as abortion or homosexual rights. These are dimensions of the multidimensional concept of ideology.

An important consideration in the choice of dimensions is their importance for a particular purpose. Jacoby (1991) offers an example. Maps MODEL the physical world. The type of map, or its dimensionality, depends on its purpose and the audience at which it is directed. A person traveling by automobile may need a map depicting roadways. This usually involves only two dimensions—latitude and longitude. A topographical engineer, however, may require more detail in the form of another dimension, height above sea level. Thus, the number of dimensions deemed important is a contextual decision.

—Edward G. Carmines and James Woods

REFERENCES

- Carmines, E. G., & McIver, J. P. (1981). *Unidimensional scaling* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–024). Beverly Hills, CA: Sage.

- Jacoby, W. G. (1991). *Data theory and dimensional analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-078). Newbury Park, CA: Sage.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Weisberg, H. F. (1974). Dimensionland: An excursion into spaces. *American Journal of Political Science*, 18, 743-776.

MULTI-ITEM MEASURES

The item is the basic unit of a psychological scale. It is a statement or question in clear, unequivocal terms about the measured characteristics (Haladyna, 1994). In social science, SCALES are used to assess people's social characteristics, such as attitudes, personalities, opinions, emotional states, personal needs, and description of their living environment.

An item is a "mini" measure that has a molecular score (Thorndike, 1967). When used in social science, multi-item measures can be superior to a single, straightforward question. There are two reasons. First, the RELIABILITY of a multi-item measure is higher than a single question. With a single question, people are less likely to give consistent answers over time. Many things can influence people's response (e.g., mood, specific thing they encountered that day). They may choose yes to a question one day and say no the other day. It is also possible that people give a wrong answer or interpret the question differently over time. On the other hand, a multi-item measure has several questions targeting the same social issue, and the final composite score is based on all questions. People are less likely to make the above mistakes to multiple items, and the composite score is more consistent over time. Thus, the multi-item measure is more reliable than a single question. Second, the VALIDITY of a multi-item measure can be higher than a single question. Many measured social characteristics are broad in scope and simply cannot be assessed with a single question. Multi-item measures will be necessary to cover more content of the measured characteristic and to fully and completely reflect the construct domain.

These issues are best illustrated with an example. To assess people's job satisfaction, a single-item measure could be as follows:

I'm not satisfied with my work. (1 = *disagree*, 2 = *slightly disagree*, 3 = *uncertain*, 4 = *slightly agree*, 5 = *agree*)

To this single question, people's responses can be inconsistent over time. Depending on their mood or specific things they encountered at work that day, they might respond very differently to this single question. Also, people may make mistakes when reading or responding. For example, they might not notice the word *not* and agree when they really disagree. Thus, this single-item measure about job satisfaction can be notoriously unreliable. Another problem is that people's feelings toward their jobs may not be simple. Job satisfaction is a very broad issue, and it includes many aspects (e.g., satisfaction with the supervisor, satisfaction with coworkers, satisfaction with work content, satisfaction with pay, etc.). Subjects may like certain aspects of their jobs but not others. The single-item measure will oversimplify people's feelings toward their jobs.

A multi-item measure can reduce the above problems. The results from a multi-item measure should be more consistent over time. With multiple items, random errors could average out (Spector, 1992). That is, with 20 items, if a respondent makes an error on 1 item, the impact on the overall score is quite minimal. More important, a multi-item measure will allow subjects to describe their feelings about different aspects of their jobs. This will greatly improve the precision and validity of the measure. Therefore, multi-item measures are one of the most important and frequently used tools in social science.

—Cong Liu

See also SCALING

REFERENCES

- Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum.
- Spector, P. (1992). *Summated rating scale construction: An introduction*. Newbury Park, CA: Sage.
- Thorndike, R. L. (1967). The analysis and selection of test items. In D. Jackson & S. Messick (Eds.), *Problems in human assessment* (pp. 201-216). New York: McGraw-Hill.

MULTILEVEL ANALYSIS

Most of statistical inference is based on replicated observations of UNITS OF ANALYSIS of one type (e.g., a sample of individuals, countries, or schools). The analysis of such observations usually is based on the

assumption that either the sampled units themselves or the corresponding RESIDUALS in some statistical model are independent and identically distributed. However, the complexity of social reality and social science theories often calls for more COMPLEX DATA SETS, which include units of analysis of more than one type. Examples are studies on educational achievement, in which pupils, teachers, classrooms, and schools might all be important units of analysis; organizational studies, with employees, departments, and firms as units of analysis; cross-national comparative research, with individuals and countries (perhaps also regions) as units of analysis; studies in GENERALIZABILITY THEORY, in which each factor defines a type of unit of analysis; and META-ANALYSIS, in which the collected research studies, the research groups that produced them, and the subjects or respondents in these studies are units of analysis and sources of unexplained variation. Frequently, but by no means always, units of analysis of different types are hierarchically nested (e.g., pupils are nested in classrooms, which, in turn, are nested in schools). *Multilevel analysis* is a general term referring to statistical methods appropriate for the analysis of data sets comprising several types of unit of analysis. The *levels* in the multilevel analysis are another name for the different types of unit of analysis. Each level of analysis will correspond to a POPULATION, so that multilevel studies will refer to several populations—in the first example, there are four populations: of pupils, teachers, classrooms, and schools. In a strictly nested data structure, the most detailed level is called the first, or the lowest, level. For example, in a data set with pupils nested in classrooms nested in schools, the pupils constitute Level 1, the classrooms Level 2, and the schools Level 3.

HIERARCHICAL LINEAR MODEL

The most important methods of multilevel analysis are variants of REGRESSION analysis designed for hierarchically nested data sets. The main model is the HIERARCHICAL LINEAR MODEL (HLM), an extension of the GENERAL LINEAR MODEL in which the probability model for the errors, or residuals, has a structure reflecting the hierarchical structure of the data. For this reason, multilevel analysis often is called hierarchical linear modeling. As an example, suppose that a researcher is studying how annual earnings of college graduates well after graduation depend on academic achievement in college. Let us assume that

the researcher collected data for a reasonable number of colleges—say, more than 30 colleges that can be regarded as a sample from a specific population of colleges, with this population being further specified to one or a few college programs—and, for each of these colleges, a random sample of the students who graduated 15 years ago. For each student, information was collected on the current income (variable Y) and the grade point average in college (denoted by the variable X in a metric where $X = 0$ is the minimum passing grade). Graduates are denoted by the letter i and colleges by j . Because graduates are nested in colleges, the numbering of graduates i may start from 1 for each college separately, and the variables are denoted by Y_{ij} and X_{ij} . The analysis could be based for college j on the model

$$Y_{ij} = a_j + b_j X_{ij} + E_{ij}.$$

This is just a linear regression model, in which the INTERCEPTS a_j and the regression coefficients b_j depend on the college and therefore are indicated with the subscript j . The fact that colleges are regarded as a random sample from a population is reflected by the assumption of random variation for the intercepts a_j and regression coefficients b_j . Denote the population mean (in the population of all colleges) of the intercepts by a and the college-specific deviations by U_{0j} , so that $a_j = a + U_{0j}$. Similarly, split the regression coefficients into the population mean and the college-specific deviations $b_j = b + U_{1j}$. Substitution of these equations then yields

$$Y_{ij} = a + bX_{ij} + U_{0j} + U_{1j}X_{ij} + E_{ij}.$$

This model has three different types of residuals: the so-called Level-1 residual E_{ij} and the Level-2 residuals U_{0j} and U_{1j} . The Level-1 residual varies over the population of graduates; the Level-2 residuals vary over the population of colleges. The residuals can be interpreted as follows. For colleges with a high value of U_{0j} , their graduates with the minimum passing grade $X = 0$ have a relatively high expected income—namely, $a + U_{0j}$. For colleges with a high value of U_{1j} , the effect of one unit GPA extra on the expected income of their graduates is relatively high—namely, $b + U_{1j}$. Graduates with a high value of E_{ij} have an income that is relatively high, given their college j and their GPA X_{ij} .

This equation is an example of the HLM; in its general form, this model can have more than one

independent variable. The first part of the equation, $a + bX_{ij}$, is called the fixed part of the model; this is a linear function of the independent variables, just like in linear regression analysis. The second part, $U_{0j} + U_{1j}X_{ij} + E_{ij}$, is called the random part and is more complicated than the random residual in linear regression analysis, as it reflects the unexplained variation E_{ij} between the graduates as well as the unexplained variation $U_{0j} + U_{1j}X_{ij}$ between the colleges. The random part of the model is what distinguishes the hierarchical linear model from the general linear model. The simplest nontrivial specification for the random part of a two-level model is a model in which only the intercept varies between Level-2 units, but the regression coefficients are the same across Level-2 units. This is called the random intercept model, and for our example, it reads

$$Y_{ij} = a + bX_{ij} + U_{0j} + E_{ij}.$$

Models in which also the regression coefficients vary randomly between Level-2 units are called random slope models (referring to graphs of the regression lines, in which the regression coefficients are the slopes of the regression lines).

The dependent variable Y in the HLM always is a variable defined at the lowest (i.e., most detailed) level of the hierarchy. An important feature of the HLM is that the independent, or explanatory, variables can be defined at any of the levels of analysis. In the example of the study of income of college graduates, suppose that the researcher is interested in the effect on earnings of alumni of college quality, as measured by college rankings, and that some meaningful college ranking score Z_j is available. In the earlier model, the college-level residuals U_{0j} and U_{1j} reflect unexplained variability between colleges. This variability could be explained partially by the college-level variable Z_j , according to the equations

$$a_j = a + c_0Z_j + U_{0j}, \quad b_j = b + c_1Z_j + U_{1j},$$

which can be regarded as linear regression equations at Level 2 for the quantities a_j and b_j , which are themselves not directly observable. Substitution of these equations into the Level-1 equation $Y_{ij} = a_j + b_jX_{ij} + E_{ij}$ yields the new model,

$$Y_{ij} = a + bX_{ij} + c_0Z_j + c_1X_{ij}Z_j + U_{0j} + U_{1j}X_{ij} + E_{ij},$$

where the parameters a and b and the residuals E_{ij} , U_{0j} , and U_{1j} now have different meanings than in the earlier model. The fixed part of this model is extended compared to the earlier model, but the random part has retained the same structure. The term $c_1X_{ij}Z_j$ in the fixed part is the INTERACTION EFFECT between the Level-1 variable X and the Level-2 variable Z . The regression coefficient c_1 expresses how much the college context (Z) modifies the effect of the individual achievement (X) on later income (Y); such an effect is called a cross-level interaction effect. The possibility of expressing how context (the “macro level”) affects relations between individual-level variables (the “micro level”) is an important reason for the popularity of multilevel modeling (see DiPrete & Forristal, 1994).

A parameter that describes the relative importance of the two levels in such a data set is the INTRACLASS CORRELATION coefficient, described in the entry with this name and also in the entry on VARIANCE COMPONENT MODELS. The similar variance ratio, when applied to residual (i.e., unexplained) variances, is called the residual intraclass correlation coefficient.

ASSUMPTIONS, ESTIMATION, AND TESTING

The standard assumptions for the HLM are the linear model expressed by the model equation, normal distributions for all residuals, and independence of the residuals for different levels and for different units in the same level. However, different residuals for the same unit, such as the random intercept U_{0j} and the random slope U_{1j} in the model above, are allowed to be correlated; they are assumed to have a multivariate normal distribution. With these assumptions, the HLM for the example above implies that outcomes for graduates of the same college are correlated due to the influences from the college—technically, due to the fact that their equations for Y_{ij} contain the same college-level residuals U_{0j} and U_{1j} . This dependence between different cases is an important departure from the assumptions of the more traditional general linear model used in regression analysis.

The parameters of the HLM can be estimated by the MAXIMUM LIKELIHOOD method. Various ALGORITHMS have been developed mainly in the 1980s (cf. Goldstein, 2003; Longford, 1993); one important algorithm is an iterative reweighted least squares algorithm (see the entry on GENERALIZED LEAST SQUARES), which alternates between estimating the regression coefficients in the fixed part and the parameters of

the random part. The regression coefficients can be tested by T -TESTS or Wald tests. The parameters defining the structure of the random part can be tested by LIKELIHOOD RATIO TESTS (also called deviance tests) or by chi-squared tests. These methods have been made available since the 1980s in dedicated multilevel software, such as HLM and MLwiN, and later also in packages that include multilevel analysis among a more general array of methods, such as M-Plus, and in some general statistical packages, such as SAS and SPSS. An overview of software capabilities is given in Goldstein (2003).

MULTIPLE LEVELS

As was illustrated already in the examples, it is not uncommon that a practical investigation involves more than two levels of analysis. In educational research, the largest contributions to achievement outcomes usually are determined by the pupil and the teacher, but the social context provided by the group of pupils in the classroom and the organizational context provided by the school, as well as the social context defined by the neighborhood, may also have important influences. In a study of academic achievement of pupils, variables defined at each of these levels of analysis could be included as explanatory variables. If there is an influence of some level of analysis, then it is to be expected that this influence will not be completely captured by the variables measured for this level of analysis, but there will be some amount of unexplained variation between the units of analysis for this level. This should then be reflected by including this unexplained variation as random residual variability in the model. The first type of residual variability is the random main effect of the units at this level, exemplified by the random intercepts U_{0j} in the two-level model above. In addition, it is possible that the effects of numerical variables (such as pupil-level variables) differ across the units of the level under consideration, which can again be modeled by random slopes such as the U_{1j} above. An important type of conclusion of analyses with multiple levels of analysis is the partitioning of unexplained variability over the various levels. This is discussed for models without random slopes in the entry on VARIANCE COMPONENT MODELS. How much unexplained variability is associated with each of the levels can provide the researcher with important directions about where to look for further explanation.

Levels of analysis can be nested or crossed. One level of analysis—the lower level—is said to be nested in another, higher level if the units of the higher level correspond to a partition into subsets of the units of the lower level (i.e., each unit of the lower level is contained in exactly one unit of the higher level). Otherwise, the levels are said to be crossed. Crossed levels of analysis often are more liable to lead to difficulties in the analysis than nested levels: Estimation algorithms may have more convergence problems, the empirical conclusions about partitioning variability over the various levels may be less clear-cut, and there may be more ambiguity in conceptual and theoretical modeling.

The use of models with multiple levels of analysis requires a sufficiently rich data set on which to base the statistical analysis. Note that for each level of analysis, the units in the data set constitute a sample from the corresponding population. Although any rule of thumb should be taken with a grain of sand, a sample size less than 20 (i.e., a level of analysis represented by less than 20 units) usually will give only quite restricted information about this population (i.e., this level of analysis), and sample sizes less than 10 should be regarded with suspicion.

LONGITUDINAL DATA

In LONGITUDINAL RESEARCH, the HLM also can be used fruitfully. In the most simple longitudinal data structure, with repeated measures on individuals, the repeated measures constitute the lower (first) and the individuals the higher (second) levels. Mostly, there will be a meaningful numerical time variable: For example, in an experimental study, this may be the time since onset of the experimental situation, and in a developmental study, this may be age. Especially for nonbalanced longitudinal data structures, in which the numbers and times of observations differ between individuals, multilevel modeling may be a natural and very convenient method. The dependence of the outcome variable on the time dimension is a crucial aspect of the model. Often, a linear dependence is a useful first approximation. This amounts to including the time of measurement as an explanatory variable; a random slope for this variable represents differential change (or growth) rates for different individuals. Often, however, dependence on time is nonlinear. In some cases, it will be possible to model this while remaining within the HLM by using several nonlinear transformations

(e.g., polynomials or splines) of time and postulating a model that is linear in these transformed time variables (see Snijders & Bosker, 1999, chap. 12). In other cases, it is better to forgo the relative simplicity of linear modeling and construct models that are not linear in the original or transformed variables or for which the Level-1 residuals are autocorrelated (cf. Verbeke & Molenberghs, 2000).

NONLINEAR MODELS

The assumption of normal distributions for the residuals is not always appropriate, although sometimes this assumption can be made more realistic by transformations of the dependent variable. In particular, for dichotomous or discrete dependent variables, other models are required. Just as the GENERALIZED LINEAR MODEL is an extension of the general linear model of regression analysis, nonlinear versions of the HLM also provide the basis of, for example, multilevel versions of LOGISTIC REGRESSION and LOGIT MODELS. These are called hierarchical generalized linear models or generalized linear mixed models (see the entry on HIERARCHICAL NONLINEAR MODELS).

—Tom A. B. Snijders

REFERENCES

- de Leeuw, J., & Kreft, I. G. G. (Eds.). (in press). *Handbook of quantitative multilevel analysis*. Dordrecht, The Netherlands: Kluwer Academic.
- DiPrete, T. A., & Forristal, J. D. (1994). Multilevel models: Methods and substance. *Annual Review of Sociology*, 20, 331–357.
- Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Hodder Arnold.
- Hox, J. J. (2002). *Multilevel analysis, techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.
- Longford, N. T. (1993). *Random coefficient models*. New York: Oxford University Press.
- Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
- Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

MULTIMETHOD-MULTITRAIT RESEARCH. See MULTIMETHOD RESEARCH

MULTIMETHOD RESEARCH

Multimethod research entails the application of two or more sources of data or research methods to the investigation of a research question or to different but highly linked research questions. Such research is also frequently referred to as *mixed methodology*. The rationale for mixed-method research is that most social research is based on findings deriving from a single research method and, as such, is vulnerable to the accusation that any findings deriving from such a study may lead to incorrect inferences and conclusions if MEASUREMENT ERROR is affecting those findings. It is rarely possible to estimate how much measurement error is having an impact on a set of findings, so that monomethod research is always suspect in this regard.

MIXED-METHOD RESEARCH AND MEASUREMENT

The rationale of mixed-method research is underpinned by the principle of TRIANGULATION, which implies that researchers should seek to ensure that they are not overreliant on a single research method and should instead employ more than one measurement procedure when investigating a research problem. Thus, the argument for mixed-method research, which in large part accounts for its growth in popularity, is that it enhances confidence in findings.

In the context of measurement considerations, mixed-method research might be envisioned in relation to different kinds of situations. One form might be that when one or more constructs that are the focus of an investigation have attracted different measurement efforts (such as different ways of measuring levels of job satisfaction), two or more approaches to measurement might be employed in combination. A second form might entail employing two or more methods of data collection. For example, in developing an approach to the examination of the nature of jobs in a firm, we might employ structured observation and structured interviews concerning apparently identical

job attributes with the same subjects. The level of agreement between the two sets of data can then be assessed.

THE MIXED-METHOD-MULTITRAIT MATRIX

One very specific context within which the term *mixed-method* is frequently encountered is in relation to the mixed-method-multitrait matrix developed by Campbell and Fiske (1959). Let us say that we have a construct that is conceptualized as comprising five dimensions or “traits” as they are referred to by Campbell and Fiske. We might develop subscales to measure each dimension and employ both structured interviewing and structured observation with respect to each trait. We can then generate a matrix that summarizes within- and between-method correlations between the different traits. The matrix allows two concerns to be addressed. First, it allows a test of the DISCRIMINANT VALIDITY of the measures because the within-method correlations should not be too large. Second, of particular importance to this entry is that the convergent validity of the between-method scores can be examined. The larger the resulting correlations, the greater the convergent validity and therefore the more confidence can be felt about the findings. Table 1 illustrates the resulting matrix.

In this matrix, a correlation of, say, $r_{i4,o2}$ indicates the correlation between trait T4, measured by a subscale administered by structured interview (*i*), and trait T2, measured by a subscale administered by structured observation (*o*). The five correlations in bold are the convergent validity correlations and indicate how well

the two approaches to the measurement of each trait (e.g., $r_{i1,o1}$) are consistent.

Care is needed over the successful establishment of convergent validity within a mixed-method research context. It is easy to assume that measurement validity has been established when convergent validity is demonstrated. However, it may be that *both* approaches to the measurement of a construct or research problem are problematic. Convergent validity enhances our confidence but of course does not eliminate completely the possibility that error remains.

COMBINING QUANTITATIVE AND QUALITATIVE RESEARCH

The account of mixed-method research so far has been firmly rooted in the tradition of measurement and of the triangulation of measurements in particular. However, the discussion of mixed-method research has increasingly been stretched to include the collection of qualitative as well as quantitative data. In other words, increasingly mixed-method research includes the combination of quantitative research and qualitative research. In this way, the discussion of mixed-method research and, indeed, of triangulation is employed not just in relation to measurement issues but also to different approaches to collecting data.

QUANTITATIVE RESEARCH and QUALITATIVE RESEARCH are sometimes taken to refer to distinct PARADIGMS and, as such, as being incompatible. Several writers have argued that quantitative and qualitative research derive from completely different epistemological and ontological traditions. They suggest that

Table 1 Mixed-Method-Multitrait Matrix

Interview	Interview					Observation				
	T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
T1	—									
T2	$r_{i2,i1}$	—								
T3	$r_{i3,i1}$	$r_{i3,i2}$	—							
T4	$r_{i4,i1}$	$r_{i4,i2}$	$r_{i4,i3}$	—						
T5	$r_{i5,i1}$	$r_{i5,i2}$	$r_{i5,i3}$	$r_{i5,i4}$	—					
Observation										
T1	$r_{o1,o1}$					—				
T2	$r_{o2,o1}$	$r_{o2,o2}$				$r_{o2,o1}$	—			
T3	$r_{o3,o1}$	$r_{o3,o2}$	$r_{o3,o3}$			$r_{o3,o1}$	$r_{o3,o2}$	—		
T4	$r_{o4,o1}$	$r_{o4,o2}$	$r_{o4,o3}$	$r_{o4,o4}$		$r_{o4,o1}$	$r_{o4,o2}$	$r_{o4,o3}$	—	
T5	$r_{o5,o1}$	$r_{o5,o2}$	$r_{o5,o3}$	$r_{o5,o4}$	$r_{o5,o5}$	$r_{o5,o1}$	$r_{o5,o2}$	$r_{o5,o3}$	$r_{o5,o4}$	—

although it is possible to employ, for example, both a structured interview (as a quintessentially quantitative research method) and ETHNOGRAPHY (as a quintessentially qualitative research method) within a single investigation, this does not and indeed cannot represent a true integration because of the divergent principles on which the two approaches are founded. The combined use of such methods represents for such authors a superficial integration of incompatible approaches.

However, most researchers have adopted a more pragmatic stance and, although recognizing the fact that quantitative and qualitative research express different epistemological and ontological commitments (such as POSITIVISM vs. INTERPRETIVISM and OBJECTIVISM vs. CONSTRUCTIONISM), they accept that much can be gained by combining their respective strengths. The apparent incommensurability of quantitative and qualitative research is usually resolved either by ignoring the epistemological and ontological issues or by asserting that research methods and sources of data are in fact much less wedded to epistemological presuppositions than is commonly supposed. With the latter argument, a research method is no more than a technique for gathering data and is largely independent of wider considerations to do with the nature of valid knowledge.

When quantitative and qualitative research are combined, it is sometimes argued that what is happening is not so much mixed-method as MULTI-STRATEGY RESEARCH (Bryman, 2001). This preference reflects a view that quantitative and qualitative research are research strategies with contrasting approaches to social inquiry and that each is associated with a cluster of research methods and research designs. These distinctions suggest the fourfold classification presented in Table 2.

Mixed-method research associated with the mixed-method-multitrait matrix essentially belongs to cell 1 in Table 2 because such an investigation combines two data collection methods associated with quantitative research. Cell 4 includes investigations that combine two or more methods associated with qualitative research. In fact, much ethnography is almost inherently mixed-method research because ethnographers frequently do more than act as participant observers. They buttress their observations with such other sources as interviewing key informants, examining documents, and engaging in nonparticipant observation.

Table 2 Monostrategy and Multistrategy Approaches to Mixed-Method Research

	<i>Quantitative Research Method</i>	<i>Qualitative Research Method</i>
Quantitative research strategy	Monostrategy multimethod research 1	Multistrategy multimethod research 2
Qualitative research strategy	Multistrategy multimethod research 3	Monostrategy multimethod research 4

APPROACHES TO COMBINING QUANTITATIVE AND QUALITATIVE RESEARCH

Cells 2 and 3 are identical in that they comprise multistrategy mixed-method research, in which at least one research method or source of data associated with both quantitative and qualitative research is employed. They can be further refined by considering the different ways that the two research strategies can be combined, an issue that has been addressed by several writers. Morgan (1998) elucidates a mechanism for classifying multistrategy research in terms of two principles: whether the quantitative or the qualitative research method is the main approach to gathering data and which research method preceded the other. This pair of distinctions yields a fourfold classification of mixed-method studies (*italics indicate the principal method*):

1. *qual* → *quant*. Examples are when a researcher conducts semistructured interviewing to develop items for a multiple-item measurement scale or when an exploratory qualitative study is carried out to generate hypotheses that can be subsequently tested by a quantitative approach.

2. *quant* → *qual*. An example is when a social survey is conducted and certain individuals are selected on the basis of characteristics highlighted in the survey for further, more intensive study using a qualitative method.

3. *quant* → *qual*. An example would be when a researcher conducts some semistructured interviewing or participant observation to help illuminate some of the factors that may be responsible for relationships between variables that have been generated from a social survey investigation.

4. *qual* → *quant*. An illustration of such a study might be when an interesting relationship between variables is discerned in an ethnographic study and a survey is then conducted to establish how far that relationship has EXTERNAL VALIDITY.

These four types of multistrategy research constitute ways in which quantitative and qualitative research can be combined and, as such, are refinements of cells 2 and 3 in Table 2. Morgan's (1998) classification is not exhaustive in that it does not consider (a) multistrategy research in which quantitative and qualitative research are conducted more or less simultaneously and (b) multistrategy research in which there is no main approach to the collection of data (i.e., they are equally weighted). Creswell (1995) distinguishes two further types of multistrategy research that reflect these two points:

5. *Parallel/simultaneous studies*. Quantitative and qualitative research methods are administered more or less at the same time.

6. *Equivalent status studies*. Two research strategies have equal status within the overall research plan.

These two types frequently co-occur, as in mixed-method research in which different research methods are employed to examine different aspects of the phenomenon being investigated. One of the more frequently encountered arguments deployed in support of using multistrategy research is that using just quantitative or qualitative research is often not sufficient to get access to all dimensions of a research question. In such a context, the quantitative and qualitative research methods are likely to be employed more or less simultaneously and to have equivalent status.

Tashakkori and Teddlie (1998) further distinguish a seventh type of multistrategy research:

7. *Designs with multilevel use of approaches*. The researcher uses "different types of methods at different levels of data aggregation" (p. 18). An example of this type of study may occur when the researcher collects different types of data for different levels of an organization (a survey of organizational members, intensive interviews with section heads, observation of work practices).

In addition, Tashakkori and Teddlie (1998) observe that the kinds of multistrategy research covered thus

far are what they call "mixed-method" studies. They observe that it is possible to distinguish these from mixed-model designs in which different phases of quantitative and qualitative research are combined. Mixed-model research occurs when, for example, a researcher collects qualitative data in an essentially exploratory manner but then submits the data to a quantitative, rather than a qualitative, data analysis.

One of the most frequently cited rationales for multistrategy research is that of seeking to establish the convergent validity of findings; in other words, this kind of rationale entails an appeal to the logic of triangulation. If quantitative findings can be confirmed with qualitative evidence (and vice versa), the credibility of the research is enhanced. Such a notion is not without problems, however. First, it is sometimes suggested that because quantitative and qualitative research derive from such contrasting epistemological and ontological positions, it is difficult to establish that one set of evidence can legitimately provide support or otherwise of the other. Second, as with all cases of triangulation, it is difficult to know how to deal with a disparity between the two sets of evidence. A common error is to assume that one is more likely to be valid than the other and to use the former as a yardstick of validity. Third, it is often suggested that the triangulation procedure is imbued with naive REALISM and that because many qualitative researchers subscribe to a constructionist position that denies the possibility of absolutely valid knowledge of the social world, it is inconsistent with much qualitative research practice. Fourth, it is easy to assume that multistrategy research is necessarily superior to monostrategy research, but this may not be so if the overall research design does not allow one research strategy to add substantially to what is known on the basis of the other research strategy. Fifth, although some social researchers are well rounded in terms of different methods and approaches, some may have strengths in one type of research rather than another, although this represents a good argument for teams with combinations of skills.

The idea of triangulation invariably presumes that the exercise is planned. However, when conducting any of the previously cited seven multistrategy research approaches, it is conceivable that the resulting quantitative and qualitative data may relate to a specific issue and therefore form an unplanned triangulation exercise. When the two sets of evidence provide convergent validity, the fact that this happens is likely to be unremarkable for the researcher, but when the two

sets of data do not converge, the researcher will need to think further about the reasons that lie behind the clash of findings. Deacon, Bryman, and Fenton (1998) report that in the course of analyzing the results from a multistrategy investigation of the representation of social research in the mass media, they found a number of disparities between their quantitative and qualitative findings. These disparities did not derive from planned triangulation exercises. For example, they found that a survey of social scientists had found a general satisfaction with media representation, but when interviewed in depth about specific instances of media accounts of their research, they were much more critical. The disparity necessitated a consideration of the different kinds and contexts of questioning and their implications for understanding the role of the media in the social sciences.

Mixed-method research is undoubtedly being used increasingly in social research. Its appeal lies in the possibilities it offers in terms of increasing the validity of an investigation. It can be employed within and across research strategies and, as such, is a flexible way of approaching research questions, although it carries certain disadvantages in terms of time and cost. Mixed-method research and multistrategy research in particular can present problems of interpretation, especially when findings are inconsistent, but the preparedness of researchers to address such problems will enhance the credibility of their work.

—Alan Bryman

REFERENCES

- Bryman, A. (2001). *Social research methods*. Oxford, UK: Oxford University Press.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Creswell, J. W. (1995). *Research design: Quantitative and qualitative approaches*. Thousand Oaks, CA: Sage.
- Deacon, D., Bryman, A., & Fenton, N. (1998). Collision or collusion? A discussion of the unplanned triangulation of quantitative and qualitative research methods. *International Journal of Social Research Methodology*, 1, 47–63.
- Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: Applications for health research. *Qualitative Health Research*, 8, 362–376.
- Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

MULTINOMIAL DISTRIBUTION

Suppose we have a RANDOM SAMPLE of N subjects, individuals, or items, and each subject is classified into exactly one of k possible categories that do not overlap in any way. That is, each subject can be classified into one and only one category. Define p_i as the probability that a subject is classified into category i for $i = 1, 2, \dots, k$, where $0 \leq p_i \leq 1$ and $\sum p_i = 1$.

The categories are then said to be mutually exclusive and exhaustive.

Define the RANDOM VARIABLE X_i as the number of subjects in the sample (the count) that are classified into category i for $i = 1, 2, \dots, k$. Obviously, we have $\sum X_i = N$. Then the joint PROBABILITY DISTRIBUTION of these k random variables is called the multinomial distribution, given by

$$f(X_1, X_2, \dots, X_k) = \binom{N}{X_1, X_2, \dots, X_k} p_1^{X_1} p_2^{X_2} \dots p_k^{X_k} \quad (1)$$

$$= \frac{N!}{X_1! X_2! \dots X_k!} p_1^{X_1} p_2^{X_2} \dots p_k^{X_k}. \quad (2)$$

The relevant PARAMETERS are $N, p_1, p_2, \dots, p_k$. There are a total of only $k - 1$ random variables involved in this distribution because of the fixed sum $\sum X_i = N$. The mean and variance of each X_i are Np_i and $Np_i(1 - p_i)$, respectively, and the covariance of the pair X_i, X_j is $-Np_i p_j$. The marginal distribution of each X_i is the BINOMIAL DISTRIBUTION with parameters N and p_i .

As an example, suppose we roll a balanced die 10 times and want to find the probability that the faces 1 and 6 each occur once and each other face occurs twice, that is, $P(X_1 = X_6 = 1, X_2 = X_3 = X_4 = X_5 = 2)$. Each $p_i = 1/6$ for a balanced die, and the calculation is 0.003751.

The term in large parentheses in (1) or, equivalently, the term involving factorials in (2) is called the multinomial coefficient.

If $k = 2$, the multinomial distribution reduces to the familiar binomial distribution with parameters N and p_1 , where the two categories are usually called success and failure, X_1 is the number of successes, and $X_2 = N - X_1$ is the number of failures.

—Jean D. Gibbons

MULTINOMIAL LOGIT

Multinomial logit (MNL) is a statistical model used to study problems with unordered categorical (nominal) dependent variables that have three or more categories. Multinomial logit estimates the PROBABILITY that each observation will be in each category of the dependent variable. MNL uses characteristics of the individual observations as independent variables, and the effects of the independent variables are allowed to vary across the categories of the dependent variable. This differs from the similar CONDITIONAL LOGIT model, which uses characteristics of the choice categories as independent variables and thus does not allow the effects of independent variables to vary across choice categories. Models that include both characteristics of the alternatives and characteristics of the individual observations are possible and are also usually referred to as conditional logits.

There are several different ways to derive the multinomial logit model. The most common approach in the social sciences is to derive the MNL as a DISCRETE choice model. Each observation is an actor selecting one alternative from a choice set of J alternatives, where J is the number of categories in the dependent variable. The utility an actor i would get from selecting alternative k from this choice set is given by

$$U_{ik} = X_i \beta_k + \varepsilon_{ik}.$$

Let y be the dependent variable with J unordered categories. Thus, the probability that an actor i would select alternative j over alternative k is

$$\begin{aligned} \Pr(y_i = k) &= \Pr(U_{ik} > U_{ij}) \\ &= \Pr(X_i \beta_k + \varepsilon_{ik} > X_i \beta_j + \varepsilon_{ij}) \\ &= \Pr(\varepsilon_{ik} - \varepsilon_{ij} > X_i \beta_j - X_i \beta_k). \end{aligned}$$

McFadden (1973) demonstrated that if the ε s are independent and identical Gumbel distributions across all observations and alternatives, this discrete choice derivation leads to a multinomial logit. MNL estimates $\Pr(y_i = k|x_i)$, where k is one of the J categories in the dependent variable. The multinomial logit model is written as

$$\Pr(y_i = k|x_i) = \frac{e^{x_i \beta_k}}{\sum_{j=1}^J e^{x_i \beta_j}}.$$

Multinomial logits are estimated using MAXIMUM LIKELIHOOD techniques. To identify the model, one

must place a constraint on the β s. Generally, the β s for one category are constrained to equal zero, although other constraints are possible (such as $\sum_{j=1}^J \beta_j = 0$). If the β s for one category are constrained to equal zero, this category becomes the baseline category, and all other β_j s indicate the effect of the independent variables on the probability of being in category j as opposed to the baseline category. Multinomial logit coefficients cannot be interpreted as REGRESSION COEFFICIENTS, and other methods (such as examining changes in $\Pr(y_i = k)$ as independent variables change) must be used to determine what effect the independent variables have on the dependent variable.

Multinomial logit is logically equivalent to estimating a series of binary logits for every pairwise comparison of categories in the dependent variable, although it is more EFFICIENT because it uses all of the data at once. MNL also has the independence of irrelevant alternatives (IIA) property, which is undesirable in many choice situations.

As an example of an application of MNL, consider the problem of estimating the effects of age, gender, and education on occupation choice, where occupation is broken down into three categories: manual labor (M), skilled labor (S), and professional (P). The results of estimating a multinomial logit for this choice problem are presented in Table 1.

Notice the coefficients vary across alternatives, and the coefficients for one alternative have been normalized to zero. This MNL tells us that as education

Table 1 Multinomial Logit Estimates, Occupation Choice (Coefficients for Manual Labor Normalized to Zero)

Independent Variables	Coefficient Value	Standard Error
<i>S/M Coefficients</i>		
Education	0.76*	0.18
Age	-0.09	0.10
Female	0.17	0.15
Constant	0.51	0.42
<i>P/M Coefficients</i>		
Education	0.89*	0.16
Age	0.21	0.15
Female	0.27*	0.13
Constant	-0.03	0.67
Number of observations	1,000	

NOTE: M = manual labor; S = skilled; P = professional.

*Statistically significant at the 95% level.

increases, individuals are less likely to select a manual labor occupation in comparison to a skilled labor or professional occupation. Women are also more likely to select a professional occupation than a manual labor occupation.

—Garrett Glasgow

REFERENCES

- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of econometrics* (pp. 105–142). New York: Academic Press.
- Train, K. (1986). *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*. Cambridge: MIT Press.

MULTINOMIAL PROBIT

REGRESSION analysis assumes a continuous (or at least interval-level) DEPENDENT VARIABLE. For many research questions in the social sciences, the CONTINUOUS VARIABLE may be unobserved and only manifested by a binomial or DICHOTOMOUS VARIABLE. For example, the unobserved variable may be the probability that a voter will choose Candidate A over Candidate B, and the observed variable is the actual choice between A and B. Analysts must choose alternative estimation procedures, such as PROBIT and LOGIT, when the continuity assumption is violated by a dichotomous dependent variable.

There are also instances when the unobserved continuous dependent variable manifests itself as three or more values or ordered categories. For example, survey respondents are often asked to indicate their reaction to a statement by responding on a 5-point scale ranging from *strongly disagree*, to *strongly agree*, or legislators may face the threefold choice of eliminating a program, funding it at the status quo, or increasing funding to a specified higher level. Variables measuring such data may be termed *multinomial*, *polychotomous*, or *n-chotomous*. Variations of standard probit and logit, called multinomial probit and MULTINOMIAL LOGIT, may be used to analyze polychotomous dependent variables.

The essence of binomial probit is to estimate the PROBABILITY of observing a 0 or 1 in the dependent variable given the values of the INDEPENDENT

VARIABLES for a particular observation. Classifying the observations is simply a matter of determining which value has the higher probability estimate. In multinomial probit, threshold parameters, commonly denoted μ_j , are estimated along with the COEFFICIENTS of the independent variables. Keeping in mind that the number of threshold parameters necessary is one less than the number of categories in the dependent variable, in the case of three categories, only two threshold parameters need to be estimated and classification would proceed as follows:

$$\begin{aligned} y &= 1 && \text{if } y^* \leq \mu_1, \\ y &= 2 && \text{if } \mu_1 \leq y^* \leq \mu_2, \\ y &= 3 && \text{if } \mu_2 \leq y^*, \end{aligned}$$

where y is the observed dependent variable and y^* is the unobserved variable. As a practical matter, μ_1 is generally normalized to be zero. Thus, the number of estimated parameters is two less than the number of categories.

Every observation in the data set has an estimated ability that it falls into each of the categories of the dependent variable. In general, the probability that an observation falls into a category is calculated as

$$\begin{aligned} \text{Prob}(y = j) &= \Phi\left(\mu_j - \sum_{i=0}^n \beta_i x_i\right) \\ &\quad - \Phi\left(\mu_{j-1} - \sum_{i=0}^n \beta_i x_i\right), \end{aligned}$$

where j is the specific category, μ_j is the threshold parameter for that category, the $\beta_i x_i$ are the estimated coefficients multiplied by the values of the corresponding n independent variables (with $\beta_0 x_0$ the constant term), and Φ is the cumulative normal probability distribution. Note that for all the probabilities to be greater than or equal to zero, $\mu_1 < \mu_2 < \mu_3 < \dots < \mu_j$ must hold (and recall that μ_1 is usually set to zero). Given the cumulative nature of the probability distribution, this can be restated more simply as $\text{Prob}(y = j) = \text{Prob}(y \leq j) - \text{Prob}(y \leq j - 1)$.

The significance of the μ_j may be tested using the familiar T -TEST. Significant and positive estimates of the μ_j confirm that the categories are ordered.

Table 1 presents sample multinomial probit results. The dependent variable, Y , is coded 1, 2, or 3. Each of the three independent variables, X_i , is coded 0 or 1. After normalizing one threshold parameter as

Table 1 Sample Results of Multinomial Probit Estimation

<i>Dependent Variable: Y</i>		
<i>Independent Variables</i>	<i>Estimated Coefficient</i>	<i>Standard Error</i>
X_1	0.90	0.29
X_2	0.55	0.26
X_3	-0.31	0.31
Constant	0.58	—
$\hat{\mu}$	0.97	0.15
N	107	
-2 Log-likelihood ratio, $df = 3$	96.23	

zero, the trichotomous dependent variable leaves only one threshold parameter to be estimated. In addition to being STATISTICALLY SIGNIFICANT, the estimated threshold parameter is positive and greater than the parameter normalized to zero.

To calculate the probabilities that a given observation falls into each of the response categories of the dependent variable, consider an observation whereby $X_1 = 1$, $X_2 = 1$, and $X_3 = 1$. From the formula above, we can use the cumulative normal probability distribution to calculate that

$$\begin{aligned}\text{Prob}(y = 1) &= \Phi\left(-\sum_{i=0}^3 \beta_i x_i\right) = .0418, \\ \text{Prob}(y = 2) &= \Phi\left(\mu - \sum_{i=0}^3 \beta_i x_i\right) \\ &\quad - \Phi\left(-\sum_{i=0}^3 \beta_i x_i\right) = .1832, \\ \text{Prob}(y = 3) &= 1 - \Phi\left(\mu - \sum_{i=0}^3 \beta_i x_i\right) = .7750.\end{aligned}$$

Notice that the calculated probabilities for the three possible categories correctly sum to 1.

As with their binary counterparts, the primary difference between the multinomial probit and multinomial logit models is the distributional function on which they are based: normal for probit, logistic for logit. The difference between the two distributions is often of little consequence unless the data set is very large or the number of observations in the tails of the

distribution is of particular importance. Interpretation of estimation results from the models is very similar, so the choice between them may lie in practical considerations such as availability of the procedure in computer programs or prior experience with one procedure or the other.

—Timothy M. Hagle

AUTHOR'S NOTE: As used here, *multinomial probit* refers to an ordered multinomial dependent variable. Some sources, however, use the term *multinomial* to refer to unordered or categorical dependent variables and simply use *ordered probit* to refer to the ordered case. Multinomial logit is the appropriate technique for estimating models with categorical dependent variables.

REFERENCES

- Liao, T. F. (1994). *Interpreting probability models: Logit, probit, and other generalized linear models*. Thousand Oaks, CA: Sage.
- Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.
- McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, 103–120.

MULTIPLE CASE STUDY

The CASE STUDY is a research method that focuses on understanding the dynamics of single settings. Although it can be used for description and deduction (Yin, 1994), our focus is on inductive theory development, an application for which the method is particularly well suited. In comparison with aggregated, statistical research, the primary advantage of case study research is its deeper understanding of specific instances of a phenomenon.

Multiple case study is a variant that includes two or more observations of the same phenomenon. This variant enables REPLICATION—that is, using multiple cases to independently confirm emerging CONSTRUCTS and propositions. It also enables extension—that is, using the cases to reveal complementary aspects of the phenomenon. The result is more robust, generalizable, and developed theory. In contrast to multiple case research, single cases may contain intriguing stories but are less likely to produce high-quality theory.

An example of the multiple case method is the inductive study by Galunic and Eisenhardt (2001) of 10 charter gains by divisions of a *Fortune* 100 high-technology firm. By studying different cases of divisions gaining new product/market responsibilities, the authors identified and described three distinct patterns for this phenomenon: new charter opportunities, charter wars, and foster homes. In addition, given the commonalities across the cases and the richness in the data, the authors were able to develop an elaborate theory of architectural innovation at the corporate level.

Multiple case research begins with data and ends with theory. Because deductive research simply reverses this order, it is not surprising that the two have similarities. These include a priori defined research questions, clearly designated POPULATION(S) from which observations are drawn, construct definition, and measurement of constructs with triangulated measures where possible. Yet, despite these similarities, there are crucial differences.

One is sampling logic. Instead of RANDOM SAMPLING of observations, THEORETICAL SAMPLING is used. That is, cases are selected to fill conceptual categories, replicate previous findings, or extend emergent theory. Furthermore, selection of cases can be adjusted during a study as it becomes clearer which cases and categories are needed to replicate or extend the emerging insights.

A second is data type. Instead of only quantitative data, both qualitative and quantitative data can be used. Indeed, the emphasis is typically on the former because qualitative data drawn from rich media such as observations, archival stories, and interviews are central to the in-depth understanding that cases provide. Qualitative data are also used because important constructs often cannot be predicted a priori. In addition, such data are invaluable for creating the underlying theoretical logic that connects constructs together into propositions.

A third difference is analysis. Instead of aggregating observations and examining them statistically, cases are treated as separate instances of the focal phenomenon, allowing replication. Because the goal is to create as close a match between data and theory, analysis involves iteratively developing tentative frameworks from one or several cases and then testing them in others (Glaser & Strauss, 1967). So, constructs are refined, patterns emerge, and relations between constructs are established. This close relationship between theory and data limits the researcher bias that

inexperienced scholars often believe is present. Once rough patterns emerge, existing literature is included in the iterations to further sharpen the emergent theory. The ultimate goal is high-quality theory.

—Filipe M. Santos and Kathleen M. Eisenhardt

REFERENCES

- Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, 14, 532–550.
- Galunic, D. C., & Eisenhardt, K. M. (2001). Architectural innovation and modular corporate forms. *Academy of Management Journal*, 44(6), 1229–1249.
- Glaser, B., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Hawthorne, NY: Aldine de Gruyter.
- Yin, R. K. (1994). *Case study research: Design and methods*. Thousand Oaks, CA: Sage.

MULTIPLE CLASSIFICATION ANALYSIS (MCA)

Also called factorial ANOVA, multiple classification analysis (MCA) is a QUANTITATIVE analysis procedure that allows the assessment of differences in subgroup means, which may have been adjusted for compositional differences in related factors and/or covariates and their effects. MCA produces the same overall results as MULTIPLE REGRESSION with DUMMY VARIABLES, although there are differences in the way the information is reported. For example, an MCA in SPSS produces an ANALYSIS OF VARIANCE with the appropriate *F* TESTS, decomposing the SUMS OF SQUARES explained by the model into the relative contributions of the factor of interest, the COVARIATE(S), and any INTERACTIONS that are specified. These *F* tests assess the ratio of the sums of squares explained by the factor(s) and covariates (if specified) adjusted by degrees of freedom ($k - 1$ categories on a k -category factor plus the number of covariates) to the unexplained sums of squares adjusted for its degrees of freedom ($N - df$ for explained sums of squares). *F* tests for each factor and for each covariate are also provided. Also reported is a table of predicted subgroup MEANS adjusted for factors and covariates as well as the DEVIATIONS (adjusted for factors and covariates) of subgroup means from the grand mean. A regression analysis in SPSS, which regresses a dependent variable on a series of DUMMY VARIABLES and covariate(s),

Table 1 Results from Multiple Classification Analysis

Marital Status	n	Predicted Mean		Deviation	
		Unadjusted	Adjusted for Covariate	Unadjusted	Adjusted for Covariate
Married	3,243	8814.4326	8737.7758	903.7891	827.1323
Widowed	124	3979.3468	4515.9674	−3931.2967	−3394.6761
Divorced	210	4696.4095	4818.3596	−3214.2340	−3092.2839
Separated	249	3257.3052	4069.9054	−4653.3383	−3840.7380
Never married	216	5087.3241	4878.8716	−2823.3194	−3035.7719

produces a comparable analysis of variance (although explained sums of squares are not decomposed) with an appropriate F test, as well as a set of regression COEFFICIENTS one can use to calculate a comparable set of subgroups means, adjusted for other factors and covariates specified in the model.

To illustrate, consider the effects of marital status and years of schooling (as the covariate) on total family income. Table 1 displays the MCA output from SPSS using the National Longitudinal Surveys cohort of mature women, who were ages 30 to 45 in 1967. The grand mean for total annual family income is \$7,910.64 (in 1967 U.S.\$). Column 1 reports marital status categories, and column 2 reports the number of cases in each subgroup. Columns 3 and 4 display the unadjusted and adjusted predicted means. The unadjusted means are the same as the subgroup means (i.e., mean family income for those in each category of marital status).

The adjusted predicted mean for each category of marital status is, for example, the predicted subgroup mean for married persons, controlling for years of schooling. The deviation column reports the comparison of the subgroup mean to the grand mean for all respondents. The unadjusted deviation indicates that the mean family income for married persons is about \$903 higher than the grand mean of \$7,911. About \$76 of that difference is due to the fact that married persons have higher than average years of schooling, and years of schooling are positively associated with total family income. Subgroup differences in education have been held to the side.

Had we estimated a regression equation with total family income as the dependent variable, specifying years of schooling and a set of dummy variables representing differences in marital status as the independent variables, we would have generated the same F -test of model fit. Furthermore, we could have used the regression coefficients to reproduce the adjusted subgroup means by substituting mean years of schooling and

the appropriate values of the dummy variables into the equation.

—Melissa A. Hardy and Chardie L. Baird

REFERENCES

- Andrews, F. M., Morgan, J. N., & Sonquist, J. A. (1967). *Multiple classification analysis: A report on a computer program for multiple regression using categorical predictors*. Ann Arbor, MI: Survey Research Center, Institute for Social Research.
- Retherford, R. D., & Choe, M. K. (1993). *Statistical models for causal analysis*. New York: John Wiley.

MULTIPLE COMPARISONS

Multiple comparisons compare $J \geq 3$ sample means from J levels of one classification variable. The ONE-WAY ANOVA, FIXED-EFFECTS MODEL simply detects the presence of significant differences in the means, not which means differ significantly.

A comparison ψ on J means is a linear combination of the following means:

$$\psi = c_1 \mu_1 + c_2 \mu_2 + \cdots + c_J \mu_J = \sum_{j=1}^J c_j \mu_j, \quad (1)$$

where the sum of the c s equals zero, but not all of the c s are zero. To estimate ψ , use

$$\hat{\psi} = c_1 \bar{Y}_1 + c_2 \bar{Y}_2 + \cdots + c_J \bar{Y}_J = \sum_{j=1}^J c_j \bar{Y}_j, \quad (2)$$

where the c s are as given for ψ . There can be many different comparisons—hence the name *multiple comparisons*. The history of tests for multiple comparisons (multiple comparison procedures [MCPs]) can be traced back to early work by Sir Ronald A. Fisher,

who suggested using individual t -tests following a significant ANOVA F test: This idea is captured and modernized in the Fisher-Hayter MCP, which is presented below. Subsequent development by Henry Scheffé and John W. Tukey, among others, led not only to their MCPs (shown below) but also to the prominence of the topic of MCPs in applied literature. After a brief coverage of basic concepts used in the area of multiple comparisons, several different MCPs are given, followed by recent developments.

BASIC CONCEPTS

A pairwise comparison has weights of 1 and -1 for two of the J means and zero for all others and thus is a difference between two means. For example, with $J = 4$, the pairwise comparison of means one and two is

$$\hat{\psi}_1 = \bar{Y}_1 - \bar{Y}_2, \quad (3)$$

with $c_1 = 1$ and $c_2 = -1$, and the other c s are zero. A nonpairwise comparison is any comparison that is not pairwise, for example,

$$\hat{\psi}_2 = \bar{Y}_1 - \frac{\bar{Y}_2 + \bar{Y}_3}{2}, \quad (4)$$

with $c_1 = 1$, $c_2 = c_3 = -1/2$, and $c_4 = 0$.

Planned (or a priori) comparisons are planned by the researcher before the results are obtained. Often, planned comparisons are based on predictions from the theory that led to the research project. Post hoc (or a posteriori) comparisons are computed after the results are obtained. With post hoc comparisons, the researcher can choose comparisons based on the results. Post hoc comparisons may or may not have a theory base for their inclusion in the study.

Orthogonality of multiple comparisons is considered for each pair of comparisons. When all the samples have a common size n , comparisons one and two are orthogonal if

$$\sum_{j=1}^J c_{1j}c_{2j} = 0. \quad (5)$$

For J means, the maximum number of comparisons that are each orthogonal to each other is $J - 1$.

If the ANOVA ASSUMPTIONS are met, the SAMPLING DISTRIBUTION of a sample comparison has known mean, known variance and known estimate of the

variance, and known shape (normal). Given this information, it is possible to form a t statistic for any given comparison,

$$t_{\hat{\psi}} = \frac{\hat{\psi} - \psi}{\sqrt{MS_W \sum_{j=1}^J \frac{c_j^2}{n_j}}}. \quad (6)$$

For equal n s and pairwise comparisons, the t statistic simplifies to

$$t = \frac{\bar{Y}_j - \bar{Y}_k}{\sqrt{\frac{MS_W(2)}{n}}}, \quad (7)$$

which has $df_W = J(n - 1) = N - J$.

If there is only a single comparison of interest, and with the usual ANOVA assumptions met, the t statistic can be used to test the hypothesis of

$$H_0 : \psi = 0, \quad (8)$$

using a critical value from the t distribution with $df = df_W$. For multiple comparisons, there are multiple hypotheses as in (8), and the t statistics must be referred to appropriate critical values for the chosen MCP. Such tests are almost always TWO-TAILED TESTS.

All of the MCPs given below are able to provide the researcher with a test of this hypothesis. However, if the researcher desires to obtain an interval estimate of ψ , then some MCPs are not appropriate because they cannot give interval estimates of the population comparisons. For those methods that can give a CONFIDENCE INTERVAL, the interval is given as

$$\begin{aligned} \hat{\psi} - \text{crit} \sqrt{\frac{MS_W}{n} \sum_{j=1}^J c_j^2} \leq \psi \leq \hat{\psi} \\ + \text{crit} \sqrt{\frac{MS_W}{n} \sum_{j=1}^J c_j^2}, \end{aligned} \quad (9)$$

where crit is an appropriate two-tailed critical value at the α level, such as t . For those researchers desiring both to test hypotheses and to obtain interval estimates, the confidence interval can be used to test the hypothesis that the comparison is zero by rejecting H_0 if the interval does not contain zero.

A researcher's primary concern should be α -control and power (see ALPHA, POWER OF THE TEST, and TYPE I ERROR). There are two basic categories of α -control

(error rates): α -control for each comparison versus α -control for some group of comparisons. The choice of category of α -control determines the value of α' to assign to each comparison and influences the power of the eventual statistic. Relative power of different MCPs can be assessed by comparing the critical values and choosing the MCP with the smallest critical value.

Setting $\alpha' = \alpha$ (typically .05) for each comparison and choosing an α -level t critical value is called error rate per comparison (ERPC). One problem with controlling α using ERPC is that p (at least one Type I error) increases as the number of comparisons (C) increases. For C orthogonal comparisons each at level α' , this is shown as $p(\text{at least one Type I error}) = 1 - (1 - \alpha')^C \leq C\alpha'$. If the C comparisons are not orthogonal, then $p(\text{at least one Type I error}) \leq 1 - (1 - \alpha')^C \leq C\alpha'$, but empirical research shows that $p(\text{at least one Type I error})$ is fairly close to $1 - (1 - \alpha')^C$.

ERPC gives a large $p(\text{at least one Type I error})$, but the power also is large. Control of error rate using ERPC gives higher power than any other method of error rate control but at the expense of high $p(\text{at least one Type I error})$. As in basic hypothesis testing, in which there is a trade-off between control of α and power, allowing high $p(\text{at least one Type I error})$ gives high power.

The second of the two basic categories of α -control is to control α for some group (family) of comparisons. Error rate per family (ERPF) is the first of two types of α -control for a family of comparisons. ERPF is not a probability, as are the other definitions of error rate. ERPF is the average number of erroneous statements (false rejections, Type I errors) made in a family of comparisons. ERPF is accomplished by setting α' at some value smaller than α , and the simplest method for doing this is to set the α' such that they add up to α . For example, if α' is the same value for all comparisons, then for C orthogonal comparisons each at level α' , ERPF is equal to $C\alpha'$. This method of controlling error rate appears in the logic of the Dunn method, as shown below.

The second type of α -control that functions for some family of comparisons is error rate familywise (ERFW). ERFW is a probability and is defined as $p(\text{at least one Type I error}) = 1 - (1 - \alpha')^C$. Even though ERFW and ERPF differ in definition and concept, in practice, if an MCP accomplishes one, it usually accomplishes the other. Thus, the basic decision a researcher faces with respect to α -control is whether

to control error rate for each comparison (ERPC) or some family of comparisons (ERPF or ERFW).

To see the relationships among these three different types of α -control, let α' be used for each of C comparisons: Using ERPC gives $\alpha = \alpha'$, using ERFW gives $\alpha < 1 - (1 - \alpha')^C$, and using ERPF gives $\alpha < C\alpha'$. Putting these together gives $\alpha' \leq 1 - (1 - \alpha')^C \leq C\alpha'$; thus, $\text{ERPC} < \text{ERFW} < \text{ERPF}$. If we set $\alpha' = .01$ for $C = 5$ orthogonal comparisons, then $\text{ERPC} = .01$, $\text{ERFW} = .049$, and $\text{ERPF} = .05$.

Types of hypotheses give configurations that exist in the population means where attention is paid to the equality, or equalities, in the means. An overall null hypothesis exists if all of the J means are equal. This is the null hypothesis tested by the overall F test—thus the name “overall null hypothesis” or “full” null hypothesis. Multiple null hypotheses exist if the overall null hypothesis is not true, but more than one subset of equal means does exist. The means must be equal within the subset, but there must be differences between the subsets. Then each of these subsets represents a null hypothesis. If there are M multiple subsets of means with equality within the subset, then there are M multiple null hypotheses. For example, if $J = 4$, and means one and two are equal but different from means three and four, which are equal, then there are $M = 2$ null hypotheses. A partial null hypothesis exists when the overall null hypothesis is not true, but some population means are equal. For $J = 4$, if the first three means are equal, but the fourth is larger than the first three, there would be a partial null hypothesis in the three equal means.

A stepwise MCP contains a test that depends in part on another test. A stepwise MCP may allow computation of comparisons only if the overall F is significant or may make testing one comparison allowable only if some other comparison is significant. MCPs that do not have such dependencies are called nonstepwise or simultaneous test procedures (STPs) because the tests may be done simultaneously.

MCPS

Many MCPs exist for the one-way ANOVA. If you are using a t statistic and have decided how to control error rate, then the choice of MCP is a choice of the critical value (theoretical distribution). Unless otherwise stated, confidence intervals can be computed from the critical value of each MCP.

Choosing α -control at ERPC gives the usual t MCP, which has the decision rule to reject H_0 if

$$|t_{\hat{\psi}}| \geq t_{df_W}^{\alpha} \quad (10)$$

and otherwise retain H_0 . Note that the usual t can be used for either pairwise or nonpairwise comparisons.

If the C comparisons are planned and α -control is chosen as ERPF (thus, ERFW), by dividing the total α into $\alpha' = \alpha/C$ for each comparison, the Dunn MCP (see BONFERRONI TECHNIQUE) is given by the decision rule to reject H_0 if

$$|t_{\hat{\psi}}| \geq t_{df_W}^{\alpha'} \quad (11)$$

and otherwise retain H_0 . The Dunn MCP can be used for either pairwise or nonpairwise comparisons.

The Tukey MCP accomplishes α -control by using an α -level critical value from the Studentized range distribution for J means and df_W , so control of α is built into the Studentized range critical value at the chosen ERFW α -level. The Tukey MCP is given by the decision rule to reject H_0 if

$$|t_{\hat{\psi}}| \geq \frac{q_{J,df_W}^{\alpha}}{\sqrt{2}} \quad (12)$$

and otherwise retain H_0 . The Tukey MCP can be used for pairwise comparisons.

The Scheffé MCP accomplishes α -control by using an α -level critical value from the F distribution for df_B and df_W , so control of α is built into the F critical value at the chosen ERFW α -level. The decision rule for Scheffé is to reject H_0 if

$$|t_{\hat{\psi}}| > \sqrt{(J-1) F_{J-1,df_W}^{\alpha}} \quad (13)$$

and otherwise retain H_0 . Literally infinitely many comparisons can be done with the Scheffé method while maintaining control of α using ERFW: pairwise comparisons, nonpairwise comparisons, orthogonal polynomials, and so on.

The Fisher-Hayter MCP allows testing of comparisons only after a significant overall ANOVA F . It has α -control at ERFW and uses a q critical value with $J-1$ as the number-of-means parameter. The decision rule is to perform the overall ANOVA F test and, if it is significant at level α , then reject H_0 if

$$|t_{\hat{\psi}}| \geq \frac{q_{J-1,df_W}^{\alpha}}{\sqrt{2}} \quad (14)$$

and otherwise retain H_0 . The Fisher-Hayter MCP is a stepwise procedure for pairwise comparisons; it does not allow computation of confidence intervals because its critical value does not take into account the first step of requiring the overall F test to be significant.

The Ryan MCP is a compilation of contributions of several statisticians. First, put the J means in order. Now consider the concept of stretch size of a comparison, the number of ordered means between and including the two means in the comparison. Stretch size is symbolized by p . The Ryan MCP has α -control at ERFW and uses a q critical value with p as the number-of-means parameter but an α that can be different depending on p . For the Ryan MCP, the decision rule is to reject H_0 if the means in the comparison are not contained in the stretch of a previously retained hypothesis and if

$$|t_{\hat{\psi}}| \geq \frac{q_{p,df_W}^{\alpha_p}}{\sqrt{2}},$$

where

$$\begin{aligned} \alpha_p &= \alpha \quad \text{for } p = J, J-1 \\ \alpha_p &= 1 - (1 - \alpha)^{p/J} \quad \text{for } p \leq J-2 \end{aligned} \quad (15)$$

and otherwise retain H_0 . The Ryan MCP is a stepwise procedure for pairwise comparisons; it does not allow computation of confidence intervals because its critical value does not take into account prior steps.

The Games-Howell (GH) MCP takes into account both unequal n s and unequal variances. It has reasonably good α -control at ERFW and uses the statistic

$$t_{jk} = \frac{\bar{Y}_j - \bar{Y}_k}{\sqrt{\frac{s_j^2}{n_j} + \frac{s_k^2}{n_k}}} \quad (16)$$

for each pair of means $j \neq k$. The decision is to reject H_0 if

$$|t_{jk}| \geq \frac{q_{J,df_{jk}}^{\alpha}}{\sqrt{2}} \quad (17)$$

and otherwise retain H_0 , where

$$df_{jk} = \frac{(s_j^2/n_j + s_k^2/n_k)^2}{\frac{(s_j^2/n_j)^2}{n_j-1} + \frac{(s_k^2/n_k)^2}{n_k-1}}. \quad (18)$$

A practical suggestion is to round df_{jk} to the nearest whole number to use as the df for the q critical value.

The GH MCP is used for pairwise comparisons. For more on basic MCPs, see Toothaker (1993). A mathematical treatment of MCPs is given by Hochberg and Tamhane (1987) and Miller (1981).

RECENT DEVELOPMENTS

Following work by Tukey, several researchers have focused on the sensitivity to extreme scores (lack of robustness) of the following statistics: the sample mean, the sample variance, and tests of means. This work examines particularly the lack of robustness when the population is a heavy-tailed distribution. Heavy-tailed departure from normality leads to inflated variance of the mean (and variance) and lack of control of α and low power for tests using means. Recent research in the area of MCPs has concentrated on using 20% trimmed means and 20% Winsorized variances to form a t -statistic. First, compute the number of scores to be trimmed from each tail as $g = [.2n]$, where the $[.2n]$ notation means to take the whole number, or integer, part of $.2n$. For each of the J groups, the steps to compute the 20% trimmed mean, $\bar{Y}_t$ include TRIMMING the g most extreme scores from each tail of the sample, summing the remaining $h = n - 2g$ scores, and dividing by h . The variance of these 20% trimmed means is estimated by a function of Winsorized variances. A Winsorized variance is computed by again focusing on the g extreme scores in each tail of each of the J samples. Let $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ represent the n ordered scores in a group. The Winsorized mean is computed as

$$\bar{Y}_w = \frac{[(g+1)Y_{(g+1)} + Y_{(g+2)} + \dots + Y_{(n-g-1)} + (g+1)Y_{(n-g)}]}{n}, \quad (19)$$

and the Winsorized variance is computed as

$$S_w^2 = \frac{(g+1)(Y_{(g+1)} - \bar{Y}_w)^2 + (Y_{(g+2)} - \bar{Y}_w)^2 + \dots + (Y_{(n-g-1)} - \bar{Y}_w)^2 + (g+1)(Y_{(n-g)} - \bar{Y}_w)^2}{n-1}. \quad (20)$$

Then we form estimates of variances of the trimmed means by using

$$d = \frac{(n-1)S_w^2}{h(h-1)} \quad (21)$$

for each of the J groups. The robust pairwise test statistic is

$$t_w = \frac{\bar{Y}_{tj} - \bar{Y}_{tk}}{\sqrt{d_j + d_k}}, \quad (22)$$

with estimated degrees of freedom of

$$df_w = \frac{(d_j + d_k)^2}{\frac{d_j^2}{h_j-1} + \frac{d_k^2}{h_k-1}}. \quad (23)$$

The robust statistic t_w with df_w can be used with a variety of MCPs, such as the Ryan MCP, above. For more on robust estimation, see Wilcox (2001). For robust estimation and subsequent MCPs, see Keselman, Lix, and Kowalchuk (1998).

—Larry E. Toothaker

REFERENCES

- Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: John Wiley.
- Keselman, H. J., Lix, L. M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods*, 3, 123–141.
- Miller, R. G. (1981). *Simultaneous statistical inference* (2nd ed.). New York: Springer-Verlag.
- Toothaker, L. E. (1993). *Multiple comparison procedures*. Newbury Park, CA: Sage.
- Wilcox, R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York: Springer.

MULTIPLE CORRELATION

The multiple correlation arises in the context of MULTIPLE REGRESSION ANALYSIS; it is a one-number summary measure of the accuracy of prediction from the regression model.

In multiple regression analysis, a single dependent variable Y (or criterion) is predicted from a set of independent variables (or predictors). In the most common form of multiple regression, multiple LINEAR REGRESSION (or ORDINARY LEAST SQUARES regression analysis), the independent variables are aggregated into a linear combination according to the following linear regression equation:

$$\hat{Y} = b_1X_1 + b_2X_2 + \dots + b_pX_p + b_0.$$

The X s are the predictors. Each predictor is multiplied by a weight, called the PARTIAL REGRESSION COEFFICIENT, $b_1, b_2, \dots, b_p$. Then, according to the regression equation, the linear combination (or weighted sum) of the scores on the set of predictors is computed. This weighted sum, noted $\hat{Y}$, is termed the *predicted score*. The regression coefficients $b_1, b_2, \dots, b_p$ are chosen in such a way that correlation between the actual dependent variable Y and the predicted score $\hat{Y}$ is as large as possible. This maximum correlation between a single criterion score Y and a linear combination of a set of X s is the multiple correlation, $R_{Y\hat{Y}}$. In usual practice, the square of this correlation is reported, referred to as the squared multiple correlation $R_{Y\hat{Y}}^2$, or *R-SQUARED*.

The squared multiple correlation is a central measure in multiple regression analysis—it summarizes the overall adequacy of the set of predictors in accounting for the dependent variable. The squared multiple correlation is the proportion of variation in the criterion that is accounted for by the set of predictors.

As an example, consider an undergraduate statistics course in which three tests are given during the semester. Suppose in a class of 50 students, we predict scores on Test 3 from scores on Tests 1 and 2 using linear regression as the method of analysis. The resulting linear regression equation is as follows:

$$\hat{Y} = .15 \text{ Test 1} + .54 \text{ Test 2} + 28.91.$$

For each student, we substitute the scores on Test 1 and Test 2 into the regression equation and compute the predicted score $\hat{Y}$ on Test 3. We then correlate these predicted scores with actual scores on Test 3; the resulting correlation is the multiple correlation. Here the multiple correlation is $R_{Y\hat{Y}} = .65$, quite a substantial correlation. Students' performance on the third test is closely related to performance on the first two tests. The squared multiple correlation $R_{Y\hat{Y}}^2 = .42$. We would describe this result by saying that 42% of the variation in the observed Test 3 scores is accounted for by scores on Test 1 and Test 2.

The multiple correlation ranges between 0 and 1. As predictors are added to the regression equation, the multiple correlation either remains the same or increases. The multiple correlation does not take into account the number of predictors. Moreover, the sample multiple correlation tends to overestimate the population multiple correlation. An adjusted squared multiple correlation that is less biased (though not

unbiased) is given as follows, where n is the number of cases and p is the number of predictors:

$$\text{Adjusted } R_{Y\hat{Y}}^2 = 1 - (1 - R_{Y\hat{Y}}^2) \times [(n - 1)/(n - p - 1)].$$

For our analysis with $n = 50$ students and $p = 2$ predictors, the adjusted $R_{Y\hat{Y}}^2 = .39$.

The multiple correlation as presented here pertains only to linear regression analysis. For other forms of regression analysis (e.g., LOGISTIC REGRESSION or POISSON REGRESSION), measures that are analogs of the multiple correlation (and squared multiple correlation) have been derived; they are not strictly multiple correlations as defined here.

—Leona S. Aiken

See also MULTIPLE REGRESSION ANALYSIS, REGRESSION ANALYSIS, R-SQUARED

REFERENCES

- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear regression models* (3rd ed.). Chicago: Irwin.

MULTIPLE CORRESPONDENCE ANALYSIS. See CORRESPONDENCE ANALYSIS

MULTIPLE REGRESSION ANALYSIS

Multiple regression analysis subsumes a broad class of statistical procedures that relate a set of INDEPENDENT VARIABLES (the predictors) to a single DEPENDENT VARIABLE (the criterion). One product of multiple regression analysis (MR) is a multiple regression equation that quantifies how each predictor relates to the criterion when all the other predictors in the equation are simultaneously taken into account. Multiple regression analysis has broad applications. It may be used descriptively, simply to summarize the relationship of a set of predictors to a criterion. It may be used for

PREDICTION. It may be used for MODEL building and theory testing.

STRUCTURE OF MULTIPLE REGRESSION

The most common form of MR (i.e., multiple LINEAR REGRESSION analysis) has the following multiple regression equation:

$$\hat{Y} = b_1X_1 + b_2X_2 + \cdots + b_pX_p + b_0.$$

The X s are the predictors. Each predictor is multiplied by a weight, called the PARTIAL REGRESSION COEFFICIENT, $b_1, b_2, \dots, b_p$. The intercept of the equation b_0 is termed the *regression constant* or *regression INTERCEPT*. In the regression equation, scores on the set of predictors are combined into a single score $\hat{Y}$, the predicted score, which is a linear combination of all the predictors.

That the independent variables and dependent variable are termed *predictors* and *criterion*, respectively, reflects the use of MR to predict individuals' scores on the criterion based on their scores on the predictors. For example, a college admissions office might develop a multiple regression equation to predict college freshman grade point average (the criterion) from three predictors: high school grade point average plus verbal and quantitative performance scores on a college admissions examination. The predicted score for each college applicant would be an estimate of the freshman college grade point average the applicant would be expected to achieve, based on the set of high school measures.

In MR, the PARTIAL REGRESSION COEFFICIENT assigned to each predictor in the regression equation takes into account the relationship of that predictor to the criterion, the relationships of all the other predictors to the criterion, and the relationships among all the predictors. The partial regression coefficient gives the unique contribution of a predictor to overall prediction over and above all the other predictors included in the equation, that is, when all other predictors are held constant.

Multiple regression analysis provides a single summary number of how accurately the set of predictors reproduces or "accounts for" the criterion, termed the *squared* MULTIPLE CORRELATION, variously noted R^2 or $R^2_{\hat{Y}\hat{Y}}$ or $R^2_{Y.12\dots p}$. The second notation, $R^2_{\hat{Y}\hat{Y}}$, indicates that the squared multiple correlation is the square of the correlation between the predicted and criterion scores. In fact, the multiple correlation is the highest

possible correlation that can be achieved between a linear combination of the predictors and the criterion. The third notation, $R^2_{Y.12\dots p}$, indicates that the correlation reflects the overall relationship of the set of predictors 1, 2, $\dots$, p to the criterion. The squared multiple correlation measures the proportion of variation in the criterion that is accounted for by the set of predictors, a measure of EFFECT SIZE for the full regression analysis.

The values of the regression coefficients are chosen to minimize the errors of prediction. For each individual, we have a measure of prediction error, the discrepancy between the individual's observed criterion score and predicted score ($Y - \hat{Y}$), termed the *residual*. In ORDINARY LEAST SQUARES regression analysis (the most common approach to ESTIMATION of regression coefficients), the partial regression coefficients are chosen so as to minimize the sum of squared residuals, the ordinary least squares (OLS) criterion. The resulting OLS regression coefficients simultaneously yield the maximum achievable multiple correlation from the set of predictors.

BUILDING A REGRESSION MODEL

A fictitious example from health psychology illustrates aspects of regression analysis and its use in model building. We wish to build a model of factors that relate to the prevention of skin cancer. The target population is young Caucasian adults living either in the Southwest or the Northeast. As the criterion, we measure their intention to protect themselves against skin cancer (Intention) on a 10-point scale. There are four predictors: (a) objective risk for skin cancer based on skin tone (Risk); (b) region of the country, Southwest or Northeast, in which the participant resides (Region); (c) participant's rating of the advantages of being tanned (Advantage); and (d) participant's perception of the injunctive norms for sun protection, that is, what he or she "should" do about sun protection (Norm). We hypothesize that Risk will be a positive predictor of Intention to sun protect, whereas Advantage of tanning will be a negative predictor. We expect higher Intention to sun protect in the Southwest with its great sun intensity. We are uncertain about the impact of injunctive Norm because young people may well not adhere to recommendations for protecting their health.

The choice of predictors illustrates the flexibility of MR. Risk, Advantage, and Norm are CONTINUOUS VARIABLES, measured on 6-point scales. Region is

CATEGORICAL. Categorical predictors require special CODING schemes to capture the fact that they are categorical and not continuous. Three coding schemes in use are dummy coding (DUMMY VARIABLE CODING), EFFECT CODING, and CONTRAST CODING. For dummy coding in the DICHOTOMOUS (two-category) case, one category is assigned the number 1, and the other is assigned 0 on a single predictor. Dummy coding is employed here with Southwest = 1 and Northeast = 0.

CORRELATIONS among the predictors and criterion for $n = 100$ cases are given in Table 1a. The final column of the correlation matrix contains the correlations of each predictor with the criterion. The expected relationships obtain. We also note SIGNIFICANT correlations among pairs of predictors. Risk is negatively correlated with Advantage. Norm is positively correlated with Region. Because Southwest is coded 1 and Northeast is coded 0, the positive correlation signifies

that the injunctive norm that one should sun protect is stronger in the Southwest.

Table 1b provides a series of bivariate regression equations (i.e., a single predictor and the criterion). There is one equation for each predictor, along with the squared multiple correlations R^2 . The regression coefficient in each single predictor equation measures the amount of change in Y for a one-unit change in the predictor. For example, Intention increases by .145 points on the 10-point scale for each 1-point increase in Risk. For the dummy-coded Region, the regression coefficient indicates that the arithmetic mean Intention is 1.049 points higher in the Southwest (coded 1) than in the Northeast (coded 0). In bivariate regression, the squared multiple correlation is simply the square of the correlation of the predictor with the criterion. We note that Risk is a statistically significant predictor of Intention. Risk alone accounts for 3.7% of the variation in

Table 1 Illustration of Multiple Regression Analysis

a. Correlations Among All Predictors and the Criterion

	<i>Risk</i>	<i>Region</i>	<i>Advantage</i>	<i>Norm</i>	<i>Intention</i>
Risk	1.000	-.075	-.203*	-.094	.193*
Region	-.075	1.000	.168+	.424*	.313*
Advantage	-.203*	.168+	1.000	.254*	-.374*
Norm	-.094	.424*	.254*	1.000	.028
Intention	.193*	.313*	-.374	.028	1.000

b. Bivariate Regression Analyses: Single-Predictor Regression Equation

1. Risk alone	$\hat{Y} = .145 \text{ Risk}^* + 3.991$	$R^2 = .037$
2. Region alone	$\hat{Y} = 1.049 \text{ Region}^* + 4.407$	$R^2 = .098$
3. Advantage alone	$\hat{Y} = -.474 \text{ Advantage}^* + 6.673$	$R^2 = .140$
4. Norm alone	$\hat{Y} = .044 \text{ Norm} + 4.760$	$R^2 = .001$

c. Prediction From Risk Plus Advantage

$$\hat{Y} = .180 \text{ Risk} - .443 \text{ Advantage}^* + 5.985 \quad R^2 = .154$$

d. Analysis of Regression: Test of the Squared Multiple Correlation, Risk Plus Advantage

<i>Sources of Variance</i>	<i>Sums of Squares</i>	<i>Degrees of Freedom</i>	<i>Mean Square</i>	<i>F Ratio</i>	<i>p</i>
Regression	43.208	2	21.606	8.858	.001
Residual	236.582	97	2.439		

e. Full Model: Prediction From All Four Predictors: Full Model

$$\hat{Y} = .202 \text{ Risk} - .516 \text{ Advantage}^* + 1.349 \text{ Region}^* - .039 \text{ Norm} + 5.680 \quad R^2 = .304$$

f. Trimmed Model and Illustration of Suppression: Prediction From Risk Plus Region

Unstandardized equation	$\hat{Y} = .319 \text{ Risk}^* + 1.104 \text{ Region}^* + 3.368$	$R^2 = .145$
Standardized equation	$\hat{Y} = .218 \text{ Risk}^* + .329 \text{ Region}^*$	$R^2 = .145$

* $p < .05$.

Y (i.e., $R^2 = .037$). Similarly, we note that Advantage is a statistically significant predictor that accounts for 14.0% of the variation in Y .

Now we use Risk and Advantage as the two predictors in a multiple regression equation, given in Table 1c. Two aspects are noteworthy. First, Risk is no longer a significant predictor of Intention as it was in the bivariate regression equation. This is so because (a) Advantage is more strongly correlated with Intention than is Risk, and (b) the two predictors are correlated ($r = -.203$). Second, together the predictors account for 15.4% of the variance (i.e., $R^2 = .154$), which is lower than the sum of the proportions of variance accounted for by each predictor taken separately, that is, $3.7\% + 14.0\% = 17.7\%$. This is due to the redundancy between Risk and Advantage in predicting Intention.

In MR, we partition (divide) the total variation in the criterion Y (typically noted SS_Y) into two nonoverlapping (orthogonal) parts. First is the predictable variation or REGRESSION SUMS OF SQUARES (typically noted $SS_{\text{regression}}$), which summarizes the extent to which the regression equation reproduces the differences among individuals on the criterion. It is literally the variation of the predicted scores around their mean. Second is the residual variation or RESIDUAL SUM OF SQUARES (noted SS_{residual}), that part of the criterion variation SS_Y that cannot be accounted for by the set of predictors. The squared multiple correlation is simply the ratio $SS_{\text{regression}}/SS_Y$, again the proportion of variation in the criterion accounted for by the set of predictors. These sources of variation are summarized into an analysis of regression summary table, illustrated in Table 1d. The two sources of variation are employed in testing the omnibus NULL HYPOTHESIS that the squared multiple correlation is zero in the population. For the numerical example, the F RATIO in the summary table leads us to reject this null hypothesis and conclude that we have accounted for a significant portion of variation in Intention from the predictors of Risk and Advantage.

Can we improve prediction by the addition of the other predictors? Hierarchical regression analysis answers this question. Hierarchical regression analysis consists of estimating a series of regression equations in which sets of predictors (one or more per set) are added sequentially to a regression equation, typically with tests of gain in the accuracy of prediction (i.e., increment in the squared multiple correlation) as each set is added. We add to our regression equation the

predictors Region and Norm, yielding the four-predictor regression equation given in Table 1e, our final model including all predictors. We see that both Advantage and Region are significant predictors of Intention in the four-predictor equation. We are not surprised that Norm has a regression coefficient of approximately zero because it is uncorrelated with Intention. We note that there is substantial gain in the squared multiple correlation from $R^2 = .154$ with only Risk and Advantage as predictors to $R^2 = .304$ with all four predictors. The increase of .150 ($.304 - .154 = .150$) is the proportion of variation in the criterion predicted by Region and Norm over and above Risk and Advantage (i.e., with Risk and Advantage partialled out or held constant). In this instance, the gain is statistically significant. The final regression equation with all four predictors is our model of Intention to protect against skin cancer. We must take care in interpretation, particularly that Risk is not a statistically significant predictor of Intention because it does correlate with Intention. We must also explain the relationship of Risk to Advantage, another important predictor of Intention.

The interplay between predictors in multiple regression analysis can lead to surprising results. Consider Table 1f, the prediction of Intention from Risk and Region. We note from the bivariate regressions that from Risk alone, $R^2 = .037$, and from Region alone, $R^2 = .098$. Yet in the equation predicting Intention from Risk and Region, $R^2 = .145$, greater than the sum from the two individual predictors. This is a special (and rare) result referred to as suppression in multiple regression. The classic explanation of a SUPPRESSOR EFFECT is that one predictor is partialing from the other predictor some irrelevant variance, thus improving the quality of the latter predictor.

The regression coefficients that we have examined are unstandardized regression coefficients. Each such coefficient is dependent on the particular units of measure of the predictor. A second set of regression coefficients, STANDARDIZED REGRESSION COEFFICIENTS, is based on standardized scores (z scores with mean of 0, standard deviation of 1) for all predictors and the criterion. Standardized regression coefficients are all measured in the same units (i.e., standard deviation units). In Table 1f, the standardized regression equation is reported for Risk and Region. The coefficient of .218 for Risk signifies that for a 1 standard deviation increase in Risk, there is a .218 standard deviation increase in Intention.

Thus far, we have only considered linear relationships of each predictor to the criterion. NON-LINEAR relationships can also be handled in multiple linear regression using either of two approaches: POLYNOMIAL regression and TRANSFORMATIONS of the predictors and/or the criterion. Polynomial regression involves creating a regression equation that has higher order functions of the predictor in the linear regression equation, an equation of the form $\hat{Y} = b_1X + b_2X^2 + b_0$ for a quadratic (one-bend) relationship or $\hat{Y} = b_1X + b_2X^2 + b_3X^3 + b_0$ for a cubic (two-bend) relationship and so forth. Suppose we expect that intention to sun protect will increase with level of habitual exposure to the sun, but only up to a point. If habitual exposure is very high, then individuals may simply ignore sun protection, so that intention to sun protect decreases at high levels of exposure. To test the hypothesis of an inverted U-shaped relationship (intention first rises and then falls as habitual sun exposure increases), we examine the b_2 coefficient for the X^2 term in a QUADRATIC EQUATION. For an inverted U-shaped relationship, this coefficient would be negative. If b_2 were essentially zero, this would indicate that the quadratic (inverted U-shaped) relationship did not hold. For a U-shaped relationship, the b_2 coefficient would be positive.

Multiple regression analysis can also treat INTERACTIONS between predictors. By *interaction*, we mean that the effect of one predictor on the criterion depends on the value of another predictor. Suppose, for example, that people in the Southwest (with relentless desert sun) were more aware of the risk of skin cancer than those in the Northeast, leading to a stronger relationship of Risk to Intention in the Southwest than in the Northeast (i.e., an interaction between Risk and Region). In multiple regression, we may specify an interaction by including an interaction term as a predictor, here between Risk and Region:

$$\hat{Y} = b_1 \text{Risk} + b_2 \text{Region} + b_3 \text{Risk} \times \text{Region} + b_0.$$

A nonzero b_3 coefficient indicates the presence of an interaction between Risk and Region (when Risk and Region are also included in the regression equation).

ASSUMPTIONS AND DIFFICULTIES

The most common linear regression model assumes that predictors are measured without error (the FIXED-EFFECTS MODEL); error in the predictors introduces bias into the estimates of regression coefficients. For

accurate statistical inference, the errors of prediction (the residuals) must follow a NORMAL DISTRIBUTION. MULTICOLLINEARITY (i.e., very high correlation among predictors) causes great instability of regression coefficients. Regression analysis is notoriously sensitive to the impact of individual errant cases (OUTLIERS); one case can markedly alter the size of a regression coefficient. REGRESSION DIAGNOSTICS assess the extent to which individual cases are INFLUENTIAL CASES that change the values of regression coefficients and, moreover, the overall fit of the regression equation.

HISTORY AND NEW EXTENSIONS

Regression analysis as a methodology made noteworthy entry into scientific literature around 1900 based on work by Galton and Pearson (see REGRESSION ANALYSIS for further exposition). The development of multiple regression as an easily implemented technique required the development of ALGORITHMS to handle numerical requirements of the analysis, particularly in the face of correlated predictors. In fact, R. A. Fisher developed the ANALYSIS OF VARIANCE with orthogonal factors due to computational difficulties in applying multiple regression analysis. With widespread availability of mainframe computers, the use of multiple regression analysis burgeoned in the 1960s and 1970s. New extensions have abounded. The GENERALIZED LINEAR MODEL subsumes methods of MR that treat a variety of dependent variables, among them LOGISTIC REGRESSION for binary and ordered category outcomes and POISSON REGRESSION for counts. Random coefficient regression and MULTILEVEL MODELING permit the analysis of clustered data with predictors measured at different levels of aggregation. Great flexibility is offered by NONPARAMETRIC REGRESSION, nonlinear regression, and ROBUST regression models.

—Leona S. Aiken

See also REGRESSION ANALYSIS

REFERENCES

- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley.
- Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear regression models* (3rd ed.). Chicago: Irwin.

MULTIPLE-INDICATOR MEASURES

MEASUREMENT of theoretical CONSTRUCTS is one of the most important steps in social research. Relating the abstract concepts described in a THEORY to empirical INDICATORS of those CONCEPTS is crucial to an unambiguous understanding of the phenomena under study. Indeed, the linkage of theoretical constructs to their empirical indicators is as important in social inquiry as the positing of the relationships between the theoretical constructs themselves.

Multiple-indicator measures refer to situations in which more than one indicator or item is used to represent a theoretical construct in contrast to a single indicator. There are several reasons why a multiple-item measure is preferable to a single item. First, many theoretical constructs are so broad and complex that they cannot be adequately represented in a single item. As a simple example, no one would argue that an individual true-false question on an American government examination is an adequate measure of the degree of knowledge of American government possessed by a student. However, if several questions concerning the subject are asked, we would get a more comprehensive assessment of the student's knowledge of the subject.

A second reason to use multiple-item measures is accuracy. Single items lack precision because they may not distinguish subtle distinctions of an attribute. In fact, if the item is dichotomous, it will only recognize two levels of the attribute.

A third reason for preferring multiple-item measures is their greater RELIABILITY. Reliability focuses on the consistency of a measure. Do repeated tests with the same instrument yield the same results? Do comparable but different tests yield the same results? Generally, multiple-indicator measures contain less RANDOM ERROR and are thus more reliable than single-item measures because random error cancels itself out across multiple measurements. For all of these reasons, multiple indicator measures usually represent theoretical constructs better than single indicators.

—Edward G. Carmines and James Woods

REFERENCES

- Carmines, E. G., & McIver, J. P. (1981). *Unidimensional scaling* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-024). Beverly Hills, CA: Sage.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Spector, P. E. (1992). *Summated rating scale construction: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-082). Newbury Park, CA: Sage.

MULTIPLICATIVE

The term *multiplicative* refers to the case when one variable is multiplied by another variable. The term is frequently used in the context of regression models, in which an outcome variable, Y , is said to be some function of the product of two variables, X and Z . A common scenario is where Y is said to be a linear function of the XZ product, such that

$$Y = \alpha + \beta XZ,$$

where XZ is the product of variable X times variable Z . A REGRESSION model is said to be “multiplicative” if it includes a multiplicative term within it. The multiplicative term can be a PREDICTOR VARIABLE multiplied by another predictor variable or a predictor variable multiplied by itself, such as

$$Y = \alpha + \beta X^2.$$

MODELS with product terms between two different variables typically imply some form of statistical INTERACTION. Models with product terms that represent squared, cubic, or higher order terms typically imply a NONLINEAR relationship between Y and X .

The fit of multiplicative models such as those described above are affected by the METRIC of the predictor variables. One can obtain a different CORRELATION between Y and XZ depending on how X and Z are scored (e.g., from -5 to $+5$ vs. from 0 to 10). There are strategies one can employ to evaluate multiplicative models so that they are invariant to such metric assumptions. These are discussed in Anderson (1981) and Jaccard and Turrisi (2003).

—James J. Jaccard

REFERENCES

- Anderson, N. H. (1981). *Methods of information integration theory*. New York: Academic Press.
- Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression*. Thousand Oaks, CA: Sage.

MULTISTAGE SAMPLING

Multistage sampling refers to SURVEY designs in which the POPULATION units are hierarchically arranged and the sample is selected in stages corresponding to the levels of the hierarchy. At each stage, only units within the higher level units selected at the previous stage are considered. (It should not be confused with *multiphase sampling*.)

The simplest type of multistage sampling is two-stage sampling. At the first stage, a sample of higher level units is selected. At the second stage, a sample of lower level units within the higher level units selected at the first stage is selected. For example, the first-stage units may be schools and the second-stage units pupils, or the first-stage units may be businesses and the second-stage units employees. In surveys of households or individuals covering a large geographical territory (e.g., national surveys), it is common for first-stage sampling units to be small geographical areas, with a modest number of addresses or households—perhaps 10 or 20—selected at the second stage within each selected first-stage unit.

Multistage designs result in samples that are “clustered” within a limited set of higher level units. If the design is to include *all* units at the latter stage—for example, all pupils in each selected school—then the sample is referred to as a CLUSTER SAMPLE because it consists of entire clusters. Otherwise, if the units are subsampled at the latter stage, the sample is said to be a *clustered sample*.

The main motivation for multistage sampling is usually cost reduction. In the case of field interview surveys, each cluster in the sample can form a workload for one interviewer. Thus, each interviewer has a number of interviews to carry out in the same location, reducing travel time relative to a single-stage (unclustered) sample. In consequence, the unit cost of an interview is reduced so a larger sample size can be achieved for a fixed budget. Often (not only for field interview surveys), there is a fixed cost associated with each first-stage unit included in the sample. For example, in a

survey of school pupils, it may be necessary to liaise with each school. The cost of this liaison, which may include visits to the school, may be independent of the number of pupils to be sampled within the school. Again, then, the unit cost of data collection can be reduced by clustering the sample of pupils within a restricted number of schools.

However, sample clustering also has a disadvantage. Because units within higher stage units (e.g., pupils within schools, households within small areas) tend to be more homogeneous than units in the population as a whole, clustering tends to reduce the precision of survey estimates. In other words, clustering tends to have the opposite effect to proportionate sample stratification (see STRATIFIED SAMPLE). Whereas stratification ensures that *all* strata are represented in the sample, clustering causes only *some* clusters to be represented in the sample. The resultant reduction in precision can be measured by the DESIGN EFFECT due to clustering. In general, the reduction in precision will be greater the larger the mean sample size per cluster and the more homogeneous the clustering units. For example, a sample of 10 pupils from each of 100 classes will result in less precision than a sample of 10 pupils from each of 100 schools if pupils within classes are more homogeneous than pupils within schools.

In designing a survey sample, the choice of clustering units and the sample size to allocate to each is important. The choice should be guided by considerations of cost and precision. Ideally, with a fixed budget, the increase in precision due to being able to afford a larger sample size should be as large as possible relative to the loss in precision due to clustering per se. The overall aim should be to maximize the accuracy of survey estimates for a fixed budget or to minimize the cost of obtaining a particular level of accuracy.

One other important aspect of multistage sampling is the control of selection probabilities. If a sample is selected in a number of conditional stages, the overall probability of selection for a particular unit is the product of the (conditional) probabilities at each stage. For example, imagine a survey of school pupils in a particular school year, with a two-stage sample design: selection of schools, followed by selection of a sample of pupils in each school. Then, the probability of selecting pupil j at school i is $P_{ij} = P_i \times P_{j|i}$, where P_i is the probability of selecting school i , and $P_{j|i}$ is the probability of selecting pupil j conditional on having selected school i . (The idea extends in an obvious way to any number of stages.)

With multistage sampling, it is important to know and to capture on the data set the selection probabilities at each stage. Only then is it possible to construct UNBIASED estimates. Furthermore, it is important to control the overall probabilities as far as possible at the design stage. In particular, in many circumstances, the most efficient design will be one that gives all population units equal selection probabilities (see DESIGN EFFECT). But there are an infinite number of possible combinations of P_i and $P_{j|i}$ that would result in the same value of P_{ij} . One obvious solution is to make each of these probabilities a constant. This is sometimes done, but it has the disadvantage that the sample size will vary across clustering units in proportion to their size. This variation could be considerable, which could cause both practical and statistical problems. A more common design is to use “probability proportional to size” (PPS) selection at each but the last stage and then select the same number of final stage units within each selected previous stage unit. In other words, $P_i = K1 \times N_i$ and $P_{j|i} = K2/N_i$, where N_i is the number of second-stage units in the i th first-stage unit (e.g., the number of pupils in the relevant year in school i , in the example of the previous paragraph), and $K1$ and $K2$ are just constant numbers. It can easily be seen that this design gives all pupils in the relevant year, regardless of which school they attend, the same overall selection probability P_{ij} while also leading to the same sample size, $K2$, of pupils in each sampled school.

—Peter Lynn

REFERENCES

- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley.
- Scheaffer, R. L., Mendenhall, W., & Ott, L. (1990). *Elementary survey sampling*. Boston: PWS-Kent.
- Stuart, A. (1984). *The ideas of sampling*. London: Griffin.

MULTISTRATEGY RESEARCH

Multistrategy research is a term employed by Layder (1993) and Bryman (2001) to refer to research that combines quantitative and qualitative research. The term has been coined to distinguish it from MULTIMETHOD RESEARCH. The latter entails the combined use of two or more methods, but these may or may not entail a combination of quantitative and qualitative

research. Multimethod research may entail the use of two or more methods within quantitative research or within qualitative research. By contrast, multistrategy research reflects the notion that quantitative and qualitative research are contrasting research strategies, each with its own set of epistemological and ontological presuppositions and criteria. Therefore, multistrategy research entails a commitment to the view that the two contrasting research strategies do not represent incommensurable PARADIGMS but instead may fruitfully be combined for a variety of reasons and in different contexts.

Multistrategy research is sometimes controversial because some writers take the view that because of their contrasting epistemological commitments, quantitative and qualitative research cannot genuinely be combined (e.g., Smith & Heshusius, 1986). According to this view, what appears to be an example of the combined use of quantitative and qualitative research (e.g., the use of both a STRUCTURED INTERVIEW-based social SURVEY with SEMISTRUCTURED INTERVIEWING) does not represent a genuine integration of the two strategies because the underlying principles are not capable of reconciliation. Instead, there is a superficial complementarity at the purely technical level. In other words, although techniques of data collection or analysis may be combined, strategies cannot. Such a view subscribes to a particular position in relation to the debate about quantitative and qualitative research—namely, one that centers on the epistemological version of the debate.

Writers and researchers who subscribe to the view that quantitative and qualitative research can be combined essentially take the position that research methods are not ineluctably embedded in epistemological and ontological presuppositions. Several different contexts for multistrategy research have been identified (Bryman, 2001). These include the following:

- using TRIANGULATION, in which the results of quantitative and qualitative research are cross-checked;
- using qualitative research to facilitate quantitative research and vice versa (e.g., when a social survey is used to help identify suitable people to be interviewed using a qualitative instrument, such as a semistructured interview);
- using one strategy to fill in the gaps that are left by the other strategy;

- using the two strategies to answer different types of research question;
- conducting qualitative research to help explain findings deriving from quantitative research; and
- gleaning an appreciation of participants' perspectives as well as addressing the concerns of the researcher.

In practice, it is difficult to separate out these different contexts of multistrategy research. For example, in their research on the mass-media representation of social science in Britain, Deacon, Bryman, and Fenton (1998) used a variety of research methods to answer different kinds of research questions. However, they found that when some of their quantitative and qualitative findings clashed, they were in fact carrying out an unplanned triangulation exercise.

—Alan Bryman

REFERENCES

- Bryman, A. (2001). *Social research methods*. Oxford, UK: Oxford University Press.
- Deacon, D., Bryman, A., & Fenton, N. (1998). Collision or collusion? A discussion of the unplanned triangulation of quantitative and qualitative research methods. *International Journal of Social Research Methodology*, 1, 47–63.
- Layder, D. (1993). *New strategies in social research*. Cambridge, UK: Polity.
- Smith, J. K., & Heshusius, L. (1986). Closing down the conversation: The end of the quantitative-qualitative debate among educational enquirers. *Educational Researcher*, 15, 4–12.

MULTIVARIATE

Statistics that deal with three or more variables at once are multivariate. For example, a MULTIPLE CORRELATION COEFFICIENT correlates two or more independent variables with a dependent variable. In MULTIPLE REGRESSION, there is a dependent variable and at least two independent variables. Any time tabular controls on a third variable are imposed in CONTINGENCY TABLES, the results are multivariate. The term is commonly juxtaposed to BIVARIATE, which looks at only two variables at once. Although the multiple variables are usually independent, they can be dependent. For example, in CANONICAL CORRELATION ANALYSIS (CCA), multiple *X* variables can be related to multiple *Y* variables. Also, an analyst may try to

explain more than one dependent variable, as in a system of SIMULTANEOUS EQUATIONS.

—Michael S. Lewis-Beck

MULTIVARIATE ANALYSIS

As the name indicates, multivariate analysis comprises a set of techniques dedicated to the analysis of data sets with more than two variables. Several of these techniques were developed recently in part because they require the computational capabilities of modern computers. Also, because most of them are recent, these techniques are not always unified in their presentation, and the choice of the proper technique for a given problem is often difficult.

This entry provides a (nonexhaustive) catalog to help researchers decide when to use a given statistical technique for a given type of data or statistical question and gives a brief description of each technique. This entry is organized according to the number of data sets to analyze: one or two (or more). With two data sets, we consider two cases: In the first case, one set of data plays the role of PREDICTORS (or INDEPENDENT) VARIABLES (IVs), and the second set of data corresponds to measurements or DEPENDENT VARIABLES (DVs). In the second case, the different sets of data correspond to different sets of DVs.

ONE DATA SET

Typically, the data tables to be analyzed are made of several measurements collected on a set of units (e.g., subjects). In general, the units are rows and the variables columns.

Interval or Ratio Level of Measurement: Principal Components Analysis (PCA)

This is the oldest and most versatile method. The goal of PCA is to decompose a data table with correlated measurements into a new set of uncorrelated (i.e., orthogonal) variables. These variables are called, depending on the context, PRINCIPAL COMPONENTS, FACTORS, EIGENVECTORS, singular vectors, or loadings. Each unit is also assigned a set of scores that correspond to its projection on the components.

The results of the analysis are often presented with graphs plotting the projections of the units onto the

components and the loadings of the variables (the so-called “circle of correlations”). The importance of each component is expressed by the variance (i.e., eigenvalue) of its projections or by the proportion of the variance explained. In this context, PCA is interpreted as an orthogonal decomposition of the variance (also called *inertia*) of a data table.

Nominal or Ordinal Level of Measurement: Correspondence Analysis (CA), Multiple Correspondence Analysis (MCA)

CA is a generalization of PCA to CONTINGENCY TABLES. The factors of CA give an orthogonal decomposition of the CHI-SQUARE associated with the table. In CA, rows and columns of the table play a symmetric role and can be represented in the same plot. When several NOMINAL variables are analyzed, CA is generalized as MCA. CA is also known as dual or OPTIMAL SCALING or reciprocal averaging.

Similarity or Distance: Multidimensional Scaling (MDS), Additive Tree, Cluster Analysis

These techniques are applied when the rows and the columns of the data table represent the same units and when the measure is a distance or a similarity. The goal of the analysis is to represent graphically these distances or similarities. MDS is used to represent the units as points on a map such that their Euclidean distances on the map approximate the original similarities (classic MDS, which is equivalent to PCA, is used for distances, and nonmetric MDS is used for similarities). Additive tree analysis and CLUSTER ANALYSIS are used to represent the units as “leaves” of a tree with the distance “on the tree” approximating the original distance or similarity.

TWO DATA SETS, CASE 1: ONE INDEPENDENT VARIABLE SET AND ONE DEPENDENT VARIABLE SET

Multiple Linear Regression Analysis (MLR)

In MLR, several IVs (which are supposed to be fixed or, equivalently, are measured without error) are used to predict with a LEAST SQUARES approach one DV. If the IVs are orthogonal, the problem reduces to a set of bivariate regressions. When the IVs are correlated, their importance is estimated from the partial coefficient of correlation. An important problem arises

when one of the IVs can be predicted from the other variables because the computations required by MLR can no longer be performed: This is called perfect multicollinearity. Some possible solutions to this problem are described in the following section.

Regression With Too Many Predictors and/or Several Dependent Variables

Partial Least Squares (PLS) Regression (PLSR)

PLSR addresses the multicollinearity problem by computing latent vectors (akin to the components of PCA), which explain both the IVs and the DVs. This very versatile technique is used when the goal is to predict more than one DV. It combines features from PCA and MLR: The score of the units as well as the loadings of the variables can be plotted as in PCA, and the DVs can be estimated (with a confidence interval) as in MLR.

Principal Components Regression (PCR)

In PCR, the IVs are first submitted to a PCA, and the scores of the units are then used as predictors in a standard MLR.

Ridge Regression (RR)

RR accommodates the multicollinearity problem by adding a small constant (the ridge) to the diagonal of the correlation matrix. This makes the computation of the estimates for MLR possible.

Reduced Rank Regression (RRR) or Redundancy Analysis

In RRR, the DVs are first submitted to a PCA and the scores of the units are then used as DVs in a series of standard MLRs in which the original IVs are used as predictors (a procedure akin to an inverse PCR).

Multivariate Analysis of Variance (MANOVA)

In MANOVA, the IVs have the same structure as in a standard ANOVA and are used to predict a set of DVs. MANOVA computes a series of ordered orthogonal linear combinations of the DVs (i.e., factors) with the constraint that the first factor generates the largest F if used in an ANOVA. The sampling distribution of this F is adjusted to take into account its construction.

Predicting a Nominal Variable: Discriminant Analysis (DA)

DA, which is mathematically equivalent to MANOVA, is used when a set of IVs is used to predict the group to which a given unit belongs (which is a nominal DV). It combines the IVs to create the largest F when the groups are used as a fixed factor in an ANOVA.

Fitting a Model: Confirmatory Factor Analysis (CFA)

In CFA, the researcher first generates one (or a few) MODEL(S) of an underlying explanatory structure (i.e., a construct), which is often expressed as a graph. Then the correlations between the DVs are fitted to this structure. Models are evaluated by comparing how well they fit the data. Variations over CFA are called STRUCTURAL EQUATION MODELING (SEM), LISREL, or EQS.

TWO (OR MORE) DATA SETS, CASE 2: TWO (OR MORE) DEPENDENT VARIABLE SETS Canonical Correlation Analysis (CC)

CC combines the DVs to find pairs of new variables (called canonical variables [CV], one for each data table) that have the highest correlation. However, the CVs, even when highly correlated, do not necessarily explain a large portion of the variance of the original tables. This makes the interpretation of the CV sometimes difficult, but CC is nonetheless an important theoretical tool because most multivariate techniques can be interpreted as a specific case of CC.

Multiple Factor Analysis (MFA)

MFA combines several data tables into one single analysis. The first step is to perform a PCA of each table. Then, each data table is normalized by dividing all the entries of the table by the first eigenvalue of its PCA. This transformation—akin to the univariate Z -score—equalizes the weight of each table in the final solution and therefore makes possible the simultaneous analysis of several heterogeneous data tables.

Multiple Correspondence Analysis (MCA)

MCA can be used to analyze several contingency tables; it generalizes CA.

Parafac and Tucker3

These techniques handle three-way data matrices by generalizing the PCA decomposition into scores and loadings to generate the three matrices of loading (one for each dimension of the data). They differ by the constraints they impose on the decomposition (Tucker3 generates orthogonal loadings, Parafac does not).

Indscal

Indscal is used when each of several subjects generates a data matrix with the same units and the same variables for all the subjects. Indscal generates a common Euclidean solution (with dimensions) and expresses the differences between subjects as differences in the importance given to the common dimensions.

Statis

Statis is used when at least one dimension of the three-way table is common to all tables (e.g., same units measured on several occasions with different variables). The first step of the method performs a PCA of each table and generates a similarity table (i.e., cross-product) between the units for each table. The similarity tables are then combined by computing a cross-product matrix and performing its PCA (without centering). The loadings on the first component of this analysis are then used as weights to compute the *compromise data table*, which is the weighted average of all the tables. The original table (and their units) is projected into the compromise space to explore their communalities and differences.

Procrustean Analysis (PA)

PA is used to compare distance tables obtained on the same objects. The first step is to represent the tables by MDS maps. Then PA finds a set of transformations that will make the position of the objects in both maps as close as possible (in the least squares sense).

—Hervé Abdi

REFERENCES

- Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling*. New York: Springer-Verlag.
- Escofier, B., & Pagès, J. (1988). *Analyses factorielles multiples* [Multiple factor analysis]. Paris: Dunod.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Upper Saddle River, NJ: Prentice Hall.

Naes, T., & Risvik, E. (Eds.). (1996). *Multivariate analysis of data in sensory science*. New York: Elsevier.

Weller, S. C., & Romney, A. K. (1990). *Metric scaling: Correspondence analysis*. Newbury Park, CA: Sage.

MULTIVARIATE ANALYSIS OF VARIANCE AND COVARIANCE (MANOVA AND MANCOVA)

The word *multi* here refers to the use of several DEPENDENT VARIABLES in a single analysis. This is in spite of the often-used term for several EXPLANATORY variables, as in MULTIPLE REGRESSION. These multiple analyses can successfully be used in fields as separate as psychology and biology—in psychological testing, each test item becomes a separate dependent variable, and in biology, there are several observations on a particular organism.

MANOVA and MANCOVA are models for the joint statistical analysis of several QUANTITATIVE dependent variables in one analysis, using the same explanatory variables for all the dependent variables. In MANOVA, all the explanatory variables are NOMINAL variables, whereas in MANCOVA, some of the explanatory variables are quantitative and some are QUALITATIVE (nominal). These models can also be extended to the regression case in which all the explanatory variables are quantitative. These three approaches are special cases of the so-called multivariate GENERAL LINEAR MODEL.

With several dependent variables and a set of explanatory variables, it is possible to separately analyze the relationship of each dependent variable to the explanatory variables, using ANALYSIS OF VARIANCE or ANALYSIS OF COVARIANCE. The difference in MANOVA and MANCOVA is that here all the dependent variables are used in the same, one analysis. This can be desirable when the several dependent variables have something in common, such as scores on different items in a psychological test. This multivariate approach also eliminates the problem of what to do for an overall significance level when many statistical tests are run on the same data.

EXPLANATION

As a simple example, suppose there are data on both the mathematics and verbal Scholastic Aptitude Test

(SAT) scores for a group of students. Are there gender and race differences in the two sets of SAT scores? It is possible to do a separate ANALYSIS OF VARIANCE for each SAT score, as expressed in the following two equations:

$$\text{Math SAT} = \text{Gender} + \text{Race} + \text{Gender} \cdot \text{Race} + \text{Residual},$$

$$\text{Verbal SAT} = \text{Gender} + \text{Race} + \text{Gender} \cdot \text{Race} + \text{Residual}.$$

Each analysis would provide the proper SUMS OF SQUARES and *P* VALUES, and it would be possible to conclude what the impacts are on the SAT scores of the two explanatory variables and their STATISTICAL INTERACTIONS.

However, two such separate analyses do not use the fact that the two explanatory variables themselves are related. If there were some way of taking this information into account, it would be possible to perform a stronger analysis and perhaps even find significant results that were not apparent in the two separate analyses. This can now be done using *one* MANOVA analysis instead of *two* separate ANOVAs.

HISTORICAL ACCOUNT

The computational complexity increases dramatically by going from analysis of variance and analysis of covariance to the corresponding multivariate analyses. Thus, the methods were not well developed and used until the necessary statistical software became available in the latter parts of the past century. Now, all major statistical software packages have the capacity to perform MANOVA and MANCOVA. Also, many of the software manuals have good explanations of the methods and how to interpret the outputs.

APPLICATIONS

Let us consider the two SAT scores and one explanatory variable, say, gender. With two separate analyses of variance, we study the distributions of the math and verbal scores separately, and we may or may not find gender differences. However, with two dependent variables, we can now create a scatterplot for the two test scores, showing their relationship. To identify the gender of each respondent, we can color the points in two different colors. It will then be clear whether the points for the two groups overlap, or the groups may show up as two very different sets of points in a

variety of possible ways. We can now see differences between the two groups that were not obvious when each variable was only considered separately. In the extreme, it is even possible to have the same mean value for both females and males on one of the two test scores. However, in the scatterplot, we may see the way in which the two groups of scores differ in major ways. Such differences would be found using one multivariate analysis of variance instead of two separate analyses.

The first step in both MANOVA and MANCOVA is to test the overall NULL HYPOTHESIS that all groups have the same means on the various dependent variables. In the example above, that implies testing that females and males have the same math scores and verbal scores and that all races have the same math scores and verbal scores. If these null hypotheses are rejected, the next step is to find *which* group means are different, just as we do in the univariate case.

Typically, there are four different ways to test the overall null hypothesis. They are known as Wilks's lambda, the Pillai-Bartlett trace, Roy's greatest characteristic root, and the Hotelling-Lawley trace. The four methods may not agree in their test of the same null hypotheses, and the choice is usually governed by issues having to do with robustness of the tests and their statistical power. Wilks's test is the most commonly used, and his test statistic has a distribution that can be approximated by an *F* DISTRIBUTION when the proper assumptions are met.

Assumptions

As with other statistical procedures, the data have to satisfy certain assumptions for the procedure to work, especially to produce meaningful *p* values. The data need to form a proper statistical random sample from an underlying population. The observations need to be obtained independently of one another. The dependent variables need to have a multivariate NORMAL DISTRIBUTION, which may be the case when each dependent variable has a normal distribution. Finally, homogeneity of variance must be the case for each dependent variable, and the correlation between any two dependent variables must be the same in all groups of observations.

—Gudmund R. Iversen

AUTHOR'S NOTE: A search on the World Wide Web of MANOVA will bring up several good explanations and examples of uses of these methods.

REFERENCES

- Bray, J. H., & Maxwell, S. E. (1985). *Multivariate analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-054). Beverly Hills, CA: Sage.
- Velleman, P. F. (1988). *DataDesk version 6.0 statistics guide*. Ithaca, NY: Data Description, Inc.

N

N (n)

The symbol(s) for SAMPLE size. Most social science research is carried out on samples from a POPULATION, rather than on the population itself. Whenever the DATA are from a sample, it is standard practice to report the sample size, such as $N = 256$, or $n = 256$. (There is no difference in practice between N and n , and each seems used about equally in the literature.) Furthermore, the reported N normally excludes cases with missing data on variables in the analysis. Thus, for example, a survey may have interviewed 1,000 respondents, but for the analysis of the six variables of interest, only 873 had no missing data, for an effective $N = 873$, not the original 1,000.

—Michael S. Lewis-Beck

N6

N6 (formerly NUD*IST), like NVivo, is a specialized computer program package for qualitative data analysis. For further information, see the website: www.scolari.co.uk/frame.html or www.scolari.co.uk/qsr/qsr.htm

—Tim Futing Liao

NARRATIVE ANALYSIS

Narrative analysis in the human sciences refers to a family of approaches to diverse kinds of texts

that have in common a storied form. As nations and governments construct preferred narratives about history, so do social movements, organizations, scientists, other professionals, ethnic/racial groups, and individuals in stories of experience. What makes such diverse texts “narrative” is sequence and consequence: Events are selected, organized, connected, and evaluated as meaningful for a particular audience. Storytellers interpret the world and experience in it; they sometimes create moral tales—how the world should be. Narratives represent storied ways of knowing and communicating (Hinchman & Hinchman, 1997). I focus here on oral narratives of personal experience.

Research interest in narrative emerged from several contemporary movements: the “narrative turn” in the human sciences away from POSITIVIST modes of inquiry and the master narratives of theory (e.g., Marxism); the “memoir boom” in literature and popular culture; identity politics in U.S., European, and transnational movements—emancipation efforts of people of color, women, gays and lesbians, and other marginalized groups; and the burgeoning therapeutic culture—exploration of personal life in therapies of various kinds. “Embedded in the lives of the ordinary, the marginalized, and the muted, personal narrative responds to the disintegration of master narratives as people make sense of experience, claim identities, and ‘get a life’ by telling and writing their stories” (Langellier, 2001, p. 700).

Among investigators, there is considerable variation in definitions of personal narrative, often linked to discipline. In social history and anthropology, narrative can refer to an entire life story, woven from the threads of interviews, observation, and documents

(e.g., Barbara Myerhoff's ethnography of elderly Jews in Venice, California). In sociolinguistics and other fields, the concept of narrative is restricted, referring to brief, topically specific stories organized around characters, setting, and plot (e.g., Labovian narratives in answer to a single interview question). In another tradition (common in psychology and sociology), personal narrative encompasses long sections of talk—extended accounts of lives in context that develop over the course of single or multiple interviews. Investigators' definitions of narrative lead to different methods of analysis, but all require them to construct texts for further analysis, that is, select and organize documents, compose FIELDNOTES, and/or choose sections of interview TRANSCRIPTS for close inspection. Narratives do not speak for themselves or have unanalyzed merit; they require interpretation when used as data in social research.

MODELS OF NARRATIVE ANALYSIS

Several typologies exist (cf. Cortazzi, 2001; Mishler, 1995). The one I sketch is an heuristic effort to describe a range of contemporary approaches particularly suited to oral narratives of personal experience (on organizational narratives, see Boje, 2001). The typology is not intended to be either hierarchical or evaluative, although I do raise questions about each. In practice, different approaches can be combined; they are not mutually exclusive, and, as with all typologies, boundaries are fuzzy. I offer several examples of each, admittedly my favorites, drawn from the field of health and illness.

Thematic Analysis

Emphasis is on the content of a text, "what" is said more than "how" it is said, the "told" rather than the "telling." A (unacknowledged) philosophy of language underpins the approach: Language is a direct and unambiguous route to meaning. As GROUNDED THEORISTS do, investigators collect many stories and inductively create conceptual groupings from the data. A typology of narratives organized by theme is the typical representational strategy, with case studies or vignettes providing illustration.

Gareth Williams (1984), in an early paper in the illness narrative genre, shows how individuals manage the assault on identity that accompanies rheumatoid arthritis by narratively reconstructing putative causes—an interpretive process that connects the body,

illness, self, and society. From analysis of how 30 individuals account for the genesis of their illness, he constructs a typology, using three cases as exemplars; they illustrate thematic variation and extend existing theory on chronic illness as biographical disruption. His interview excerpts often take the classic, temporally ordered narrative form, but analysis of formal properties is not attempted.

Carole Cain (1991) goes a bit further in her study of identity acquisition among members of an Alcoholics Anonymous group, in which she uses observation and interviews. There are common propositions about drinking in the classic AA story, which new members acquire as they participate in the organization; over time, they learn to place the events and experiences in their lives into a patterned life story that is recognizable to AA audiences. She identifies a general cultural story and analyzes how it shapes the "personal" stories of group members—key moments in the drinking career, often told as episodes. By examining narrative structure in a beginning way, her work segues into the text type.

The thematic approach is useful for theorizing across a number of cases—finding common thematic elements across research participants and the events they report. A typology can be constructed to elaborate a developing theory. Because interest lies in the content of speech, analysts interpret what is said by focusing on the meaning that any competent user of the language would find in a story. Language is viewed as a resource, not a topic of investigation. But does the approach mimic OBJECTIVIST modes of inquiry, suggesting themes are unmediated by the investigator's theoretical perspective, interests, and mode of questioning? The contexts of an utterance—in the interview, in wider institutional and cultural discourses—are not usually studied. Readers must assume that when many narratives are grouped into a similar thematic category, everyone in the group means the same thing by what he or she says. What happens to ambiguities, "deviant" responses that don't fit into a typology, the unspoken?

Structural Analysis

Emphasis shifts to the telling, the *way* a story is told. Although thematic content does not slip away, focus is equally on form—how a teller, by selecting particular narrative devices, makes a story persuasive. Unlike the thematic approach, language is treated seriously—an object for close investigation—over and beyond its referential content.

Arguably the first method of narrative analysis developed by William Labov and colleagues more than 30 years ago, this structural approach analyzes the *function* of a clause in the overall narrative—the communicative work it accomplishes. Labov (1982) later modified the approach to examine first-person accounts of violence—brief, topically centered, and temporally ordered stories—but he retained the basic components of a narrative’s structure: the abstract (summary and/or point of the story); orientation (to time, place, characters, and situation); complicating action (the event sequence, or plot, usually with a crisis and turning point); evaluation (where the narrator steps back from the action to comment on meaning and communicate emotion—the “soul” of the narrative); resolution (the outcome of the plot); and a coda (ending the story and bringing action back to the present). Not all stories contain all elements, and they can occur in varying sequences. Labov’s microanalysis convincingly shows how violent actions (in bars, on the street, etc.) are the outcome of speech acts gone awry. From a small corpus of narratives and the prior work of Goffman, he develops a theory of the rules of requests that explains violent eruptions in various settings experienced by a diverse group of narrators.

An ethnopoetic structural approach is suitable for lengthy narratives that do not take the classic temporal story form. Building on the work of Dell Hymes and others, James Gee (1991) analyzes the speech of a woman hospitalized with schizophrenia and finds it artful and meaningful. Episodically (rather than temporally) organized, the narrative is parsed into idea units, stanzas, strophes, and parts based on how the narrative is *spoken*. Meaning and interpretation are constrained by features of the spoken narrative. Gee develops a theory of units of discourse that goes beyond the sentential.

Because structural approaches require examination of syntactic and prosodic features of talk, they are not suitable for large numbers, but they can be very useful for detailed case studies and comparison of several narrative accounts. Microanalysis of a few cases can build theories that relate language and meaning in ways that are missed when transparency is assumed, as in thematic analysis. Investigators must decide, depending on the focus of a project, how much transcription detail is necessary. There is the danger that interview excerpts can become unreadable for those unfamiliar with sociolinguistics, compromising communication across disciplinary boundaries.

Like the thematic approach, strict application of the structural approach can decontextualize narratives by ignoring historical, interactional, and institutional factors. Research settings and relationships constrain what can be narrated and shape the way a particular story develops.

Interactional Analysis

Here, the emphasis is on the dialogic process between teller and listener. Narratives of experience are occasioned in particular settings, such as medical, social service, and court situations, where storyteller and questioner jointly participate in conversation. Attention to thematic content and narrative structure are not abandoned in the interactional approach, but interest shifts to storytelling as a process of coconstruction, where teller and listener create meaning collaboratively. Stories of personal experience, organized around the life-world of the teller, may be inserted into question-and-answer exchanges. The approach requires transcripts that include all participants in the conversation, and it is strengthened when paralinguistic features of interaction are included as well.

Some research questions require interactional analysis. Jack Clark and Elliot Mishler (1992) sought to distinguish the features that differentiated “attentive” medical interviews from others. By analyzing pauses, interruptions, topic chaining, and other aspects of conversation, they show how medical interviews can (and cannot) result in patient narratives that provide knowledge for accurate diagnosis and treatment.

Susan Bell (1999) compares the illness narratives of two women, separated in time by the women’s health movement and activism. Interrogating her participation in the research interviews, she shows how the emergent narratives are situated historically and politically. Contexts shape possibilities in the women’s lives, their experiences of illness, and the specific illness narratives the women produce collaboratively with the author. Microanalysis of language and interaction, in addition to narrative organization and structure, are essential to her method.

An interactional approach is useful for studies of relationships between speakers in diverse field settings (courts of law, classrooms, social service organizations, psychotherapy offices, and the research interview itself). Like structural approaches, studies of interaction typically represent speech in all its

complexity, not simply as a vehicle for content. As in CONVERSATION ANALYSIS, transcripts may be difficult for the uninitiated. Pauses, disfluencies, and other aspects of talk are typically included, but what cannot be represented in a transcript (unlike a videotape) is the unspoken. What happens to gesture, gaze, and other displays that are enacted and embodied?

Performative Analysis

Extending the interactional approach, interest goes beyond the spoken word, and, as the stage metaphor implies, storytelling is seen as performance by a “self” with a past who involves, persuades, and (perhaps) moves an audience through language and gesture, “doing” rather than telling alone. Variation exists in the performative approach, ranging from dramaturgic to narrative as praxis—a form of social action. Consequently, narrative researchers may analyze different features: actors allowed on stage in an oral narrative (e.g., characters and their positionings in a story, including narrator/protagonist); settings (the conditions of performance and setting of the story performed); the enactment of dialogue between characters (reported speech); and audience response (the listener[s] who interprets the drama as it unfolds, and the interpreter in later reading[s]). Performative analysis is emergent in narrative studies, although the dramaturgic view originated with Goffman, and researchers are experimenting with it in studies of identities—vested presentations of “self” (Riessman, 2003).

Kristin Langellier and Eric Peterson (2003) provide a compelling theory and many empirical examples, ranging from detailed analysis of family (group) storytelling to an illness narrative told by a breast cancer survivor. They analyze the positioning of storyteller, audience, and characters in each performance; storytelling is a communicative practice that is embodied, situated, material, discursive, and open to legitimation and critique.

The performative view is appropriate for studies of communication practices and for detailed studies of identity construction—how narrators want to be known and precisely how they involve the audience in “doing” their identities. The approach invites study of how audiences are implicated in the art of narrative performance. As Wolfgang Iser and reader-response theorists suggest, readers are the ultimate interpreters, perhaps reading a narrative differently from either teller or investigator. Integrating the visual (through filming and

photography) with the spoken narrative represents an innovative contemporary turn (Radley & Taylor, 2003).

CONCLUSION

Analysis of narrative is no longer the province of literary study alone; it has penetrated all of the human sciences and practicing professions. The various methods reviewed are suited to different kinds of projects and texts, but each provides a way to systematically study personal narratives of experience. Critics argue (legitimately, in some cases) that narrative research can reify the interior “self,” pretend to offer an “authentic” voice—unalloyed subjective truth—and idealize individual agency (Atkinson & Silverman, 1997; Bury, 2001). There is a real danger of overpersonalizing the personal narrative.

Narrative approaches are not appropriate for studies of large numbers of nameless and faceless subjects. Some modes of analysis are slow and painstaking, requiring attention to subtlety: nuances of speech, the organization of a response, relations between researcher and subject, social and historical contexts—cultural narratives that make “personal” stories possible. In a recent reflexive turn, scholars in AUTOETHNOGRAPHY and other traditions are producing their own narratives, relating their biographies to their research materials (Riessman, 2002).

Narratives do not mirror the past, they refract it. Imagination and strategic interests influence how storytellers choose to connect events and make them meaningful for others. Narratives are useful in research precisely because storytellers interpret the past rather than reproduce it as it was. The “truths” of narrative accounts are not in their faithful representations of a past world but in the shifting connections they forge among past, present, and future. They offer storytellers a way to reimagine lives (as narratives do for nations, organizations, and ethnic/racial and other groups forming collective identities). Building on C. Wright Mills, narrative analysis can forge connections between personal biography and social structure—the personal and the political.

—Catherine Kohler Riessman

REFERENCES

- Atkinson, P., & Silverman, D. (1997). Kundera’s “immortality”: The interview society and the invention of self. *Qualitative Inquiry*, 3(3), 304–325.

- Bell, S. E. (1999). Narratives and lives: Women's health politics and the diagnosis of cancer for DES daughters. *Narrative Inquiry*, 9(2), 1–43.
- Boje, D. M. (2001). *Narrative methods for organizational and communication research*. Thousand Oaks, CA: Sage.
- Bury, M. (2001). Illness narratives: Fact or fiction? *Sociology of Health and Illness*, 23(3), 263–285.
- Cain, C. (1991). Personal stories: Identity acquisition and self-understanding in Alcoholics Anonymous. *Ethos*, 19, 210–253.
- Clark, J. A., & Mishler, E. G. (1992). Attending to patients' stories: Reframing the clinical task. *Sociology of Health and Illness*, 14, 344–370.
- Cortazzi, M. (2001). Narrative analysis in ethnography. In P. Atkinson, A. Coffey, S. Delamont, J. Lofland, & L. Lofland (Eds.), *Handbook of ethnography*. Thousand Oaks, CA: Sage.
- Gee, J. P. (1991). A linguistic approach to narrative. *Journal of Narrative and Life History*, 1, 15–39.
- Hinchman, L. P., & Hinchman, S. K. (Eds.). (1997). *Memory, identity, community: The idea of narrative in the human sciences*. Albany: State University of New York Press.
- Labov, W. (1982). Speech actions and reactions in personal narrative. In D. Tannen (Ed.), *Analyzing discourse: Text and talk*. Washington, DC: Georgetown University Press.
- Langellier, K. M. (2001). Personal narrative. In M. Jolly (Ed.), *Encyclopedia of life writing: Autobiographical and biographical forms* (Vol. 2). London: Fitzroy Dearborn.
- Langellier, K. M., & Peterson, E. E. (2003). *Performing narrative: The communicative practice of storytelling*. Philadelphia: Temple University Press.
- Mishler, E. G. (1995). Models of narrative analysis: A typology. *Journal of Narrative and Life History*, 5(2), 87–123.
- Radley, A., & Taylor, D. (2003). Remembering one's stay in hospital: A study in recovery, photography and forgetting. *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine*, 7(2), 129–159.
- Riessman, C. K. (2002). Doing justice: Positioning the interpreter in narrative work. In W. Patterson (Ed.), *Strategic narrative: New perspectives on the power of personal and cultural storytelling* (pp. 195–216). Lanham, MA: Lexington Books.
- Riessman, C. K. (2003). Performing identities in illness narrative: Masculinity and multiple sclerosis. *Qualitative Research*, 3(1), 5–33.
- Williams, G. (1984). The genesis of chronic illness: Narrative re-construction. *Sociology of Health & Illness*, 6(2), 175–200.

forms, ranging from tightly bounded ones that recount specific past events (with clear beginnings, middles, and ends) to narratives that traverse temporal and geographical space—biographical accounts that refer to entire lives or careers.

The idea of narrative interviewing represents a major shift in perspective in the human sciences about the research interview itself. The question-and-answer (stimulus/response) model gives way to viewing the interview as a discursive accomplishment. Participants engage in an evolving conversation; narrator and listener/questioner, collaboratively, produce and make meaning of events and experiences that the narrator reports (Mishler, 1986). The “facilitating” interviewer and the vessel-like “respondent” are replaced by two active participants who jointly produce meaning (Gubrium & Holstein, 2002). Narrative interviewing has more in common with contemporary ETHNOGRAPHY than with mainstream social science interviewing practice that relies on discrete OPEN-ENDED QUESTIONS and/or CLOSED-ENDED QUESTIONS.

ENCOURAGING NARRATION

When the interview is viewed as a conversation—a discourse between speakers—rules of everyday conversation apply: turn taking; relevancy; and entrance and exit talk to transition into, and return from, a story world. One story can lead to another; as narrator and questioner/listener negotiate spaces for these extended turns, it helps to explore associations and meanings that might connect several stories. If we want to learn about experience in all its complexity, details count: specific incidents, not general evaluations of experience. Narrative accounts require longer turns at talk than are typical in “natural” conversation, certainly in mainstream research practice.

Opening up the research interview to extended narration by a research participant requires investigators to give up some control. Although we have particular experiential paths we want to cover, narrative interviewing means following participants down their trails. Genuine discoveries about a phenomenon can come from power sharing in interviews.

Narratives can emerge at the most unexpected times, even in answer to fixed-response (yes/no) questions (Riessman, 2002). But certain kinds of open-ended questions are more likely than others to provide narrative opportunities. Compare “When did X happen?” which requests a discrete piece of

NARRATIVE INTERVIEWING

How does narrative interviewing differ from the classic in-depth interview? The first term suggests the generation of detailed “stories” of experience, not generalized descriptions. But narratives come in many

information, with “Tell me what happened . . . and then what happened?” which asks for an extended account of some past time. Some investigators, after introductions, invite a participant to tell their story—how an illness began, for example. But experience always exceeds its description and narrativization; events may be fleetingly summarized and given little significance. Only with further questioning can participants recall the details, turning points, and other shifts in cognition, emotion, and action. In my own research on disruptions in the expected life course, such as divorce, I have posed the question: “Can you remember a particular time when . . .?” I might probe further: “What happened that makes you remember that particular moment in your marriage?” Cortazzi and colleagues (Cortazzi, Jin, Wall, & Cavendish, 2001), studying the education of health professionals, asked: “Have you had a breakthrough in your learning recently?” “Oh yes” typically followed, and then a long narrative with an outpouring of emotion and metaphor about a breakthrough—“a clap of thunder,” as one student said.

In general, less structure in interview schedules gives greater control to research participants—interviewee and interviewer alike—to jointly construct narratives using available cultural forms. Not all parents tell stories to children on a routine basis, and not all cultures are orally based (in some groups, of course, stories are the primary way to communicate about the past). Storytelling as a way of knowing and telling is differentially invoked by participants in research interviews. Not all narratives are “stories” in the strict (sociolinguistic) sense of the term.

Collecting and comparing different forms of telling about the past can be fruitful. A graduate student in my class, Janice Goodman, studied a group of Sudanese boys that had traversed many countries before the boys were finally accepted into the United States as “legitimate” refugees. Their narrative accounts differed from one another in significant ways, but all contained particular events in the narrative sequence—for example, crossing a river filled with crocodiles in which many boys perished. The absence of emotion and evaluation in the narratives contrasts with accounts of other refugee groups, reflecting perhaps the young age of my student’s participants, their relationship to her, and cultural and psychological factors.

Sometimes, it is next to impossible for a participant to narrate experience in spoken language alone. Wendy Luttrell (2003), working as an ethnographer in a classroom for pregnant teens, most of whom were African

American, expected “stories” from each girl about key events: learning of her pregnancy, telling her mother and boyfriend, making the decision to keep the baby, and other moments. She confronted silence instead, only to discover a world of narrative as she encouraged the girls’ artistic productions and role-plays. When she asked them to discuss their artwork, they performed narratives about the key moments for each other—group storytelling. It is a limitation for investigators to rely only on the texts we have constructed from individual interviews, our “holy transcripts.” Innovations among contemporary scholars include combining observation, sustained relationships and conversations over time with participants, even visual data with narrative interviewing (e.g., videotaping of participants’ environments, photographs they take, and their responses to photographs of others).

In sum, narrative interviewing is not a set of techniques, nor is it necessarily natural. If used creatively in some research situations, it offers a way for us to forge dialogic relationships and greater communicative equality in social research.

—Catherine Kohler Riessman

REFERENCES

- Cortazzi, M., Jin, L., Wall, D., & Cavendish, S. (2001). Sharing learning through narrative communication. *International Journal of Language and Communication Disorders*, 36, 252–257.
- Gubrium, J. F., & Holstein, J. A. (2002). From the individual interview to the interview society. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research: Context and method*. Thousand Oaks, CA: Sage.
- Luttrell, W. (2003). *Pregnant bodies, fertile minds: Gender, race, and the schooling of pregnant teens*. New York: Routledge.
- Mishler, E. G. (1986). *Research interviewing: Context and narrative*. Cambridge, MA: Harvard University Press.
- Riessman, C. K. (2002). Positioning gender identity in narratives of infertility: South Indian women’s lives in context. In M. C. Inhorn & F. van Balen (Eds.), *Infertility around the globe: New thinking on childlessness, gender, and reproductive technologies*. Berkeley: University of California Press.

NATIVE RESEARCH

The predominance of POSITIVISM in EPISTEMOLOGY and science has long emphasized the

necessity for researchers to conduct their work with OBJECTIVITY and methodologies through which concerns about VALIDITY, RELIABILITY, and generalizability are obviated. The phrase “GOING NATIVE” is often attributed to Bronislaw Malinowski based upon his extensive experience studying the Trobriand Islanders of New Guinea (Malinowski, 1922). In his reflections upon the relationship between the anthropologist and “subjects” in the field, Malinowski suggested that, contrary to the tradition of distancing oneself from objects of study research, one should instead “grasp the native’s point of view, his relations to life, to realize *his* vision of *his* world” (p. 290). That is, scientists in a foreign milieu should go native, engaging in the research endeavor in a more interactive and reflexive manner with those peoples and cultures they study.

Over the seven decades since Malinowski first suggested the notion of going native, researchers across disciplines have focused upon the ways they might gain access to or study those populations and/or topics that are dissimilar to the researcher by gender, race, class, nationality, or other characteristics. Over the past 10 years, however, the field of anthropology has been primarily responsible for coining the nomenclature of the “native,” “indigenous,” or “insider” researcher (Hayano, 1979; Kanuha, 2000; Malinowski, 1922; Narayan, 1993; Ohnuki-Tierney, 1984; Reed-Danahay, 1997) in which ethnographers study social or identity groups of which they are members and/or social problem areas about which they have direct and often personal experience. Examples of native research would be female researchers studying breast cancer, ethnographers with HIV/AIDS conducting focus groups with HIV-positive clients, or Hiroshima-born scientists surveying Japanese survivors of the atomic bomb.

The benefits of researchers conducting studies by/about themselves include access to previously unstudied or hard-to-reach populations, the nature and depth of “insider” knowledge that accentuates the possibilities and levels of analytical interpretation, and a complex understanding of methods and data that can be born only of LIVED EXPERIENCE in a specific cultural context. However, native research also includes challenges. Balancing the multiple identities of being at once a PARTICIPANT OBSERVER *and* researcher requires unique skills and understandings to distance emotionally and intellectually from data in order to enhance analyses without predisposition. This is particularly difficult when the native researcher’s primary social

identities often mirror those he or she is studying. In addition, an essential criterion for the native researcher is to have not only some prior knowledge of the population, but also the ability to be accepted as a member of one’s own group (an insider) even while conducting oneself as a researcher (an outsider).

The conundrums of native research require sensitivity to those native researchers who bring both “insider” perspectives and “outsider” methods for studying them.

—Valli Kalei Kanuha

REFERENCES

- Hayano, D. (1979). Auto-ethnography: Paradigms, problems and prospects. *Human Organization*, 38(1), 99–104.
- Kanuha, V. (2000). “Being” native vs. “going native”: The challenge of doing research as an insider. *Social Work*, 45(5), 439–447.
- Malinowski, B. (1922). *Argonauts of the Western Pacific*. New York: Holt, Rinehart and Winston.
- Narayan, K. (1993). How native is the “native” anthropologist? *American Anthropologist*, 95, 671–686.
- Ohnuki-Tierney, E. (1984). “Native” anthropologists. *American Ethnologist*, 11, 584–586.
- Reed-Danahay, D. E. (Ed.). (1997). *Auto/ethnography: Rewriting the self and the social*. Oxford, UK: Berg.

NATURAL EXPERIMENT

The natural or naturalistic experiment is “natural” in the sense that it makes investigative use of real-life, naturally occurring happenings as they unfold, without the imposition of any CONTROL or manipulation on the part of the researcher(s), and usually without any preconceived notions on what the research outcomes will be. In the social sciences, empirical inquiries of this kind make use of natural settings such as playgrounds, schools, workplaces, neighborhoods, communities, and people’s homes. With their naturally occurring events, these contexts provide scope for social scientists to unobtrusively gather and then analyze experimental data making use of QUANTITATIVE RESEARCH and/or QUALITATIVE RESEARCH methods (e.g., IN-DEPTH INTERVIEWS, CASE STUDIES, OBSERVATIONS, QUESTIONNAIRES, ATTITUDE SCALES, ratings, test scores). Experiments of this kind study the effect of natural phenomena on the response/DEPENDENT VARIABLE.

The naturalistic experiment is often contrasted with LABORATORY EXPERIMENTS or FIELD EXPERIMENTS, where events under investigation are intentionally and systematically manipulated by the researcher. For example, the researcher introduces, withdraws, or manipulates some variables while holding others constant. But despite these differences, the two methodologies can complement each other, with the natural experiment helping to determine whether or not the outcomes of carefully controlled laboratory studies occur in real life (i.e., establishing EXTERNAL VALIDITY). On the other hand, indicative results may prompt more rigorous experimental work. Another merit of the natural experiment is that certain research may be undertaken only by using this approach, because experimental research would be unethical, too costly, or too difficult.

Potential weaknesses include problems to do with SAMPLING, limitations in generalizing results to wider populations, and problems in extrapolating causal inferences from data. Although sampling can be randomized in experimental research, it may be outside the researchers' control in the naturalistic experiment. Thus, researchers may be obliged to depend upon self-selection. In the St. Helena study (see Charlton, Gunter, & Hannan, 2002), for example, comparisons were made between viewers and nonviewers. Problems about whether findings can be generalized to other places, times, and people are also linked to sampling. Findings can be generalized only when there is evidence to suggest that the sample under scrutiny is a REPRESENTATIVE SAMPLE of the general population. Finally, causal inferences can be difficult to make unless controls are in place to show that other factors have not influenced results. For example, Cook, Campbell, and Peracchio (1990) talk of four general kinds of natural experimental research design, only one of which allows—but does not guarantee—causal inferences.

The study by Charlton et al. (2002) is an example of the naturalistic experiment. Researchers monitored children's behavior before and then for 5 years after the availability of broadcast television using a MULTIMETHOD research approach (observations, ratings, self-reports, questionnaires, FOCUS GROUP discussions, DIARIES, and CONTENT ANALYSIS) to take advantage of a naturally occurring event in order to assess television's impact upon viewers.

—Tony Charlton

REFERENCES

- Charlton, T., Gunter, B., & Hannan, A. (2002). *Broadcast television effects in a remote community*. Hillsdale, NJ: Lawrence Erlbaum.
- Cook, T. D., Campbell, D. T., & Peracchio, L. (1990). Quasi-experimentation. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organisational psychology* (2nd ed., Vol. 1, pp. 491–576). Chicago: Rand McNally.

NATURALISM

The term is used in three different ways. *Philosophical naturalism* is the assertion that there is an objective, natural order in the world to which all things belong (see Papineau, 1993). It is, then, a denial of Cartesian dualism, which assumes a separation of physical things and mind. Philosophical naturalism usually implies materialism, that what we experience as mind can be accounted for materially (although it would be just as consistent with a naturalistic position to say all was reducible to mind). Philosophical naturalism is no more than metaphysical reasoning or speculation, but it underlies the most important version of naturalism, *scientific naturalism*.

SCIENTIFIC NATURALISM

Scientific naturalism spans a wide range of views. Some maintain that a unity of nature implies a unity of approach to investigating nature, and this is likely to entail the best methods of science currently available to us. At the other extreme is a much harder version of naturalism that embraces an extreme form of physicalism and reductionism. In this view, all phenomena are reducible to physical properties, and things such as the senses and moral values are simply epiphenomena.

Although natural scientists will place themselves at different points on such a continuum, almost by definition they are naturalists of some kind, but this is not always the case with those who study the social world and even call themselves social “scientists.” If the ontological assumptions of naturalism are applied to the social world, then it must follow that the latter is continuous with, or arises from, the physical world (Williams, 2000, p. 49). Furthermore, if we are to explain the relation of humans to the world in terms of that order, with appropriate adaptations, scientific methods can be used to study the social world. This

view underlies two of the most prominent approaches in social science, POSITIVISM and REALISM.

What kind of assumptions we can make about the world and the methods that follow from these assumptions divide realists in both the natural and the social sciences, but they are united in the belief that those things considered the proper outcomes of natural science—explanations and generalizations—are appropriate for social science. Naturalists do not deny that there are important differences in the way the social world (as compared to the physical world) manifests itself to other human beings, and it is accepted that this will produce methodological differences. However, it is argued, such differences already exist in natural science. Physicists, for example, use experimental methods, whereas astronomers use passive observational methods.

OBJECTIONS TO NATURALISM

There are a number of objections to naturalism from within both philosophy and social studies.

Antirationalists, such as postmodernists and those in the social studies of science, direct their attack not at the ontological claims of naturalism, but at science itself, claiming that it is “just one kind of conversation or one form of social organisation” (Kincaid, 1996, p. 8) and can claim no special epistemological privilege. This renders naturalism redundant, because it has no means by which it can demonstrate the existence of a natural order. However, the counterargument must come not from a defense of naturalism, but from scientific method itself (Williams, 2000, chap. 4).

A second argument against scientific naturalism is that it is tautological. Science is the only way we can know the natural order, and what counts as natural are those things discovered by science. Again, this can be answered only obliquely by maintaining that naturalism does not depend on any particular historical form science might take, only that form of science that provides the best approximation to the truth about puzzles of nature or society.

The most important objections to naturalism as a basis for study of the *social* world are methodological. Most of these are agnostic as to whether there is a “unity of nature,” but instead focus on the way the social world manifests itself as a product of consciousness. It is said that the resulting feedback mechanisms produce an indeterminacy that renders

impossible the identification of the regularities necessary for explanation and generalization. Therefore, studies of the social world cannot be scientific in the way that, say, physics or biology are. This claim leads to two broad methodological positions, although the first is something of an antimethodology: (a) science is the most reliable way to discover truths about the world, but unfortunately, social science is unable to be scientific, and it follows, therefore, that there can be no reliable route to knowledge of the social world; (b) the social world can be investigated, but it must be studied from a humanistic perspective that places emphasis on understanding rather than explanation.

There are a number of responses to the above methodological objections. The first is that naturalism and methodology are relatively independent of each other, because the natural sciences and different social sciences, or even research problems, will require different methods. For the most part, anthropology would not be well served by an explanatory survey, and, conversely, political polling would be impossible with ethnographic methods. Second, explanations and generalizations are just as possible in many areas of the social world as they are in any other complex system. They are likely to be probabilistic, but this is the case in many studies of the physical world (e.g., turbulence, genetic mutation, etc.). Third, some aspects of the social world will defy our efforts to know about them, but this is equally true in the physical world.

The debate about naturalism in social science is made difficult by its conflation with positivism. Critics of the latter position in social science have seen its particular conception of science as standing in for all scientific study of the social world. This conflation is problematic partly because it is misleading, but also because it ignores that although natural scientists are naturalists, they are often antipositivists.

NATURALISM IN ETHNOGRAPHY

The third use of the term *naturalism* is found in ethnography. This usage has nothing to do with the previous, older employment of the term and is indeed commonly used by humanists opposed to scientific naturalism. What might be termed *ethnographic naturalism* is the view that people create their own social realities from meanings, and to know those social realities, the researcher should minimize the

amount of interference he or she makes in that reality while conducting research. This position is principally associated with SYMBOLIC INTERACTIONISM.

—Malcolm Williams

REFERENCES

- Kincaid, H. (1996). *Philosophical foundations of the social sciences: Analyzing controversies in social research*. Cambridge, UK: Cambridge University Press.
- Papineau, D. (1993). *Philosophical naturalism*. Oxford, UK: Basil Blackwell.
- Williams, M. (2000). *Science and social science: An introduction*. London: Routledge.

NATURALISTIC INQUIRY

Research designs are naturalistic to the extent that the research takes place in real-world settings and the researcher does not attempt to manipulate the phenomenon of interest (e.g., a group, event, program, community, relationship, or interaction). The phenomenon of interest unfolds “naturally” in that it has no predetermined course established by and for the researcher such as would occur in a laboratory or other controlled setting. Observations take place in real-world settings, and people are interviewed with open-ended questions in places and under conditions that are comfortable for and familiar to them.

Egon Guba's (1978) classic treatise on naturalistic inquiry identified two dimensions along which types of scientific inquiry can be described: (a) the extent to which the scientist manipulates some phenomenon in advance in order to study it; and (b) the extent to which constraints are placed on outputs, that is, the extent to which *predetermined* categories or variables are used to describe the phenomenon under study. He then defined naturalistic inquiry as a discovery-oriented approach that minimizes investigator manipulation of the study setting and places no prior constraints on what the outcomes of the research will be. Naturalistic inquiry contrasts with controlled EXPERIMENTAL DESIGNS, where, ideally, the investigator controls study conditions by manipulating, changing, or holding constant external influences and where a limited set of outcome variables is measured. Open-ended, conversation-like interviews as a form of naturalistic inquiry contrast with questionnaires that have predetermined response categories.

In the simplest form of controlled experimental inquiry, the researcher gathers data at two points in time, pretest and posttest, and compares the treatment group to some control group on a limited set of standardized measures. Such designs assume a single, identifiable, isolated, and measurable treatment. Moreover, such designs assume that, once introduced, the treatment remains relatively constant and unchanging.

In program evaluation, for example, controlled experimental evaluation designs work best when it is possible to limit program adaptation and improvement so as not to interfere with the rigor of the research design. In contrast, under real-world conditions where programs are subject to change and redirection, naturalistic inquiry replaces the fixed treatment/outcome emphasis of the controlled experiment with a dynamic process orientation that documents actual operations and impacts of a process, program, or intervention over a period of time. The evaluator sets out to understand and document the day-to-day reality of participants in the program, making no attempt to manipulate, control, or eliminate situational variables or program developments, but accepts the complexity of a changing program reality. The data of the evaluation include whatever emerges as important to understanding what participants experience.

Natural experiments occur when the observer is present during a real-world change to document a phenomenon before and after the change. Natural experiments can involve comparing two groups, one of which experiences some change while the other does not. What makes such studies naturalistic is that real-world participants direct the changes, not the researcher, as in the laboratory.

However, the distinction is not as simple as being in the field versus being in the laboratory; rather, the degree to which a design is naturalistic falls along a continuum with completely open fieldwork on one end and completely controlled laboratory conditions on the other end, but with varying degrees of researcher control and manipulation between these endpoints. For example, the very presence of a researcher asking questions can be an intervention that reduces the natural unfolding of events. Or, in the case of a program evaluation, providing feedback to staff can be an intervention that reduces the natural unfolding of events. *Unobtrusive observations* are used as an inquiry strategy when the inquirer wants to minimize data collection as an intervention.

Alternative approaches to agricultural research illuminate the differences between controlled and naturalistic inquiries. In agricultural testing stations, field experiments are conducted in which researchers carefully vary an intervention with predetermined measures and rigorous controls, as in crop fertilizer studies. In contrast, naturalistic inquiry in agricultural research involves observing, documenting, and studying farmers' own "experiments" on their own farms, undertaken without researcher direction or interference.

Anthropologists engage in naturalistic inquiry when they do ethnographic fieldwork. Sociologists use naturalistic inquiry to do PARTICIPANT OBSERVATION studies in neighborhoods and organizations. Program evaluators employ naturalistic inquiry to observe how a program is being implemented by staff. Naturalistic inquiry is used by social scientists generally to study how social phenomena and human interactions unfold in real-world settings.

EMERGENT DESIGN FLEXIBILITY

Naturalistic inquiry designs cannot be completely specified in advance of fieldwork. Although the design will specify an initial focus, plans for observations, and initial guiding interview questions, the naturalistic and inductive nature of the inquiry makes it both impossible and inappropriate to specify operational variables, state testable hypotheses, finalize either instrumentation, or completely predetermine sampling approaches. A naturalistic design unfolds or emerges as fieldwork unfolds (Patton, 2002). Design flexibility stems from the open-ended nature of naturalistic inquiry.

—Michael Quinn Patton

REFERENCES

- Guba, E. G. (1978). *Toward a methodology of naturalistic inquiry in educational evaluation*. CSE Monograph Series in Evaluation, No. 8. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.
- Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks, CA: Sage.

on, as well as the probabilities associated with those values. A random variable, Y , has a negative binomial probability distribution if its probability mass function (pmf) follows the form:

$$p(y) = \binom{y-1}{r-1} p^r q^{y-r},$$

$$y = r, r+1, r+2, \dots, 0 \leq p \leq 1,$$

where p is the probability of success (occurrence of an event), q is the probability of failure (no occurrence of event), and r is the first success (occurrence). The data follow from a process that can have only two outcomes: success or failure (occurrence or non-occurrence). A random variable following a negative binomial distribution has the following mean and variance:

$$\mu = E(Y) = \left(\frac{r}{p}\right) \quad \text{and} \quad \sigma^2 = V(Y) = \frac{r(1-p)}{p^2}.$$

The type of REGRESSION model and underlying probability distribution on which you decide to base your estimates depends on the characteristics of the data. The principles that are used to derive the distribution parallel the principles (characteristics) of the data. When your data are of the event count form, certain characteristics are implicit, namely, that there is a lower bound of zero and that there is a finite number your observations may take. Such data cannot follow a NORMAL DISTRIBUTION. The POISSON REGRESSION model is often appropriate for such data.

However, in some situations, two fundamental assumptions of the Poisson do not hold, namely, that the data are independent and identically distributed. The negative binomial distribution generalizes the Poisson by relaxing these assumptions, which makes it a better distribution on which to base your estimates if you have reason to believe your observations are not independent or occurring constantly. To test whether the Poisson or the negative binomial is appropriate, it is necessary to check for under- and overdispersion of the Poisson regression. Overdispersion is present if the sample mean is larger than the variance of the dependent event count variable. The effect of overdispersion on standard errors and significance statistics is similar to HETEROSKEDASTICITY in OLS estimation.

The negative binomial regression model is useful if, for example, you are studying participation in protests. It is likely that one person's decision to participate is not entirely independent from another's; some members of

NEGATIVE BINOMIAL DISTRIBUTION

A PROBABILITY distribution is the full range of conceivable values that a RANDOM VARIABLE could take

a student political organization may be more prone to take part than those who are not affiliated with any political organizations. This is a situation of violating the Poisson assumption of independence. The second assumption is violated if you, say, studied the establishment of left-wing social movements over time. If the government of a country shifted from authoritarian to democratic, it is plausible that you would find an inconstant rate of occurrence—social movement development would be more restricted under the repressive regime than under the new government. This is obviously not restricted to the TIME SERIES case; the rate of occurrence may not, for example, be constant across states in cross-sectional analysis.

If you decide that the negative binomial regression model is appropriate for your data, keep in mind that you will have to identify the distribution of the rate of change (if you suspect inconsistency). This has the potential to be very difficult, and MISSPECIFICATION will lead to inaccurate PARAMETER ESTIMATES. If you do not have a theoretical basis on which to project a λ , you may consider using a generalized event count model.

—Christine Mahoney

REFERENCES

- Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge, UK: Cambridge University Press.
- King, G. (1998). *Unifying political methodology: The likelihood theory of statistical inference*. Ann Arbor: University of Michigan Press.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

NEGATIVE CASE

In qualitative analysis, data are usually grouped to form patterns (identified as constructs) with the expectation that there will be some degree of VARIATION within those patterns. However, either through the process of purposeful searching or by happenstance, it is possible to come across a case that does not fit within the pattern, however broadly the construct is defined. This case is usually referred to as a “negative case” because it seems contrary to the general pattern. For example, suppose one was studying caregivers of people with Alzheimer’s disease, and one of

the constructs identified from the study is “emotional distress” (Khurana, 1995). All of the caregivers who are interviewed express some degree of emotional distress. However, as data gathering continues, the researcher happens upon a participant who expresses little, if any, distress. Coming across this one participant does not necessarily negate the general pattern, but the case does provide additional insight into the construct of “emotional distress.”

Negative cases can serve as comparative cases enabling analysts to explain more clearly what a CONCEPT is and what it is not in terms of its properties (i.e., identify the boundaries of a concept). Negative cases can help delineate important conditions pertaining to a concept, for instance, enabling the researcher to explain why some people who are caregivers are more or less likely to develop emotional distress. They also enable analysts to extend or modify the original construct, possibly adding to its explanatory power. Negative cases can also lend a degree of VALIDITY to a study by demonstrating that the analyst is willing to consider alternatives and has indeed searched for other possible explanations (Miles & Huberman, 1994). Because social science research deals with human beings, one can expect that there will always be exceptions to every identified pattern (Patton, 1990).

Does one always discover negative cases when conducting research? The answer is that whether or not one discovers the negative case depends on how hard and how long one looks. At the same time, there are no set rules for how long one should continue the search for the “negative case” (Glaser & Strauss, 1967). There are always constraints on a research project in terms of time, money, and access to a population. However, a researcher should make a concerted effort to find the exceptions to the pattern and include reference to the search process when writing up the research; then, when and if such a case is discovered, he or she should bring the additional insights and alternative explanations into the discussion of findings. This will not only add to the credibility to the research but also enhance understanding of the phenomenon under investigation, benefiting participants and users of the research.

—Juliet M. Corbin

REFERENCES

- Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Chicago: Aldine.

- Khurana, B. (1995). *The older spouse caregiver: Paradox and pain of Alzheimer's disease*. Unpublished doctoral dissertation, Center for Psychological Studies, Albany, CA.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

NESTED DESIGN

Nested design is a research design in which levels of one factor (say, Factor *B*) are hierarchically subsumed under (or nested within) levels of another factor (say, Factor *A*). As a result, assessing the complete combination of *A* and *B* levels is not possible in a nested design. For example, one might wish to study the effect of Internet search strategies (Factor *A*) on college students' information-seeking efficiency (the dependent variable). Six classes of freshmen English at a state college are randomly selected; three classes are assigned to the "linear search" condition and the other three to the "nonlinear search" condition. Table 1 illustrates the research design.

Because freshmen enrolled in these classes form "intact groups," they cannot be randomly assigned to the two treatment conditions on an individual basis. Furthermore, their learning processes and behaviors are likely to be mutually dependent; differences in students' information-seeking behavior among classes are embedded within each treatment condition. This restriction makes this design a nested design rather than a fully crossed design, and the nested design is denoted as *B*(*A*).

The statistical model assumed for the above nested design is

$$Y_{ijk} = \mu + \alpha_j + \beta_{k(j)} + \varepsilon_{i(jk)},$$

$$(i = 1, \dots, n; j = 1, \dots, p,$$

$$\text{and } k = 1, \dots, q),$$

where

Y_{ijk} is the score of the i th observation in the j th level of Factor *A* and k th level of Factor *B*.

μ is the grand mean, therefore, a constant, of the population of observations.

α_j is the effect of the j th treatment condition of Factor *A*; algebraically, it equals the deviation of the population mean ($\mu_{.j}$) from the grand mean (μ). It is a constant for all observations' dependent score in the j th condition, subject to the restriction that all α_j sum to zero across all treatment conditions.

$\beta_{k(j)}$ is the effect for the k th treatment condition of Factor *B*, nested within the j th level of Factor *A*; algebraically, it equals the deviation of the population mean ($\mu_{jk.}$) in the k th and j th combined level from the grand mean (μ). It is a constant for all observations' dependent score in the k th condition, nested within Factor *A*'s j th condition. The effect is assumed to be normally distributed in its underlying population.

$\varepsilon_{i(jk)}$ is the random error effect associated with the i th observation in the j th condition of Factor *A* and k th condition of Factor *B*. It is a random variable that is normally distributed in the underlying population and is independent of $\beta_{k(j)}$.

Table 1

		Factor A Internet Search Strategy		Row Mean
		Treatment 1 Linear Search Strategy	Treatment 2 Nonlinear Search Strategy	
Factor B Freshman English Class	Class 1	$\bar{Y}_{1(1)}$		$\bar{Y}_{.1}$
	Class 2	$\bar{Y}_{2(1)}$		$\bar{Y}_{.2}$
	Class 3	$\bar{Y}_{3(1)}$		$\bar{Y}_{.3}$
	Class 4		$\bar{Y}_{4(2)}$	$\bar{Y}_{.4}$
	Class 5		$\bar{Y}_{5(2)}$	$\bar{Y}_{.5}$
	Class 6		$\bar{Y}_{6(2)}$	$\bar{Y}_{.6}$
	Column mean	$\bar{Y}_{1.}$	$\bar{Y}_{2.}$	Grand mean = $\bar{Y}_{...}$

Table 2

		Factor A Internet Search Strategy			
		Factor C Prior Experience With Internet	Treatment 1 Linear Search Strategy	Treatment 2 Nonlinear Search Strategy	Row Mean
Factor B Freshman English Class	Class 1	High	$\bar{Y}_{1(1)1}$		$\bar{Y}_{.1.}$
		Low	$\bar{Y}_{1(1)2}$		
	Class 2	High	$\bar{Y}_{2(1)1}$		$\bar{Y}_{.2.}$
		Low	$\bar{Y}_{2(1)2}$		
	Class 3	High	$\bar{Y}_{3(1)1}$		$\bar{Y}_{.3.}$
		Low	$\bar{Y}_{3(1)2}$		
	Class 4	High		$\bar{Y}_{4(2)1}$	$\bar{Y}_{.4.}$
		Low		$\bar{Y}_{4(2)2}$	
	Class 5	High		$\bar{Y}_{5(2)1}$	$\bar{Y}_{.5.}$
		Low		$\bar{Y}_{5(2)2}$	
	Class 6	High		$\bar{Y}_{6(2)1}$	$\bar{Y}_{.6.}$
		Low		$\bar{Y}_{6(2)2}$	
Column mean			$\bar{Y}_{1..}$	$\bar{Y}_{2..}$	Grand mean = $\bar{Y}_{..}$

The example represents a balanced, completely randomized nested design. Each cell (i.e., the combination of *A* and *B* levels) has $n = 4$ observations; thus, the design is balanced. It is also completely randomized because observations are assigned randomly to levels (or conditions) of *A*, and no blocking variable is involved in this example. Additionally, if the researcher wishes to control for individual differences on a blocking variable such as “prior experience with Internet search,” the design suitable for his or her research is the completely randomized block nested design. Table 2 explains the layout of this design and its difference from the completely randomized nested design.

Variations of the balanced, completely randomized nested design include the following: unbalanced nested design, balanced or unbalanced randomized block nested design, balanced or unbalanced randomized factorial nested design, and balanced or unbalanced randomized block factorial nested design. For details on these designs, readers should consult Kirk (1995) and Maxwell and Delaney (1990).

Nested design is alternatively called hierarchical design; it is used most often in QUASI-EXPERIMENTAL

studies in which researchers have little or no control over random assignment of observations into treatment conditions. The design is popular, and sometimes necessary, among curriculum studies and clinical, sociological, and ethological research in which observations (units of analysis) belong to intact groups (such as classes, therapeutic groups, gangs, cages); these intact groups cannot be dismantled to allow for a random assignment of observations into different treatment conditions. Because of the nonindependence among units of observations in nested designs, methodologists have recommended using group means (i.e., class means in the example above) to perform statistical analyses. This recommendation proves to be restrictive, unnecessary, and less powerful than alternatives derived directly from individual data and proper statistical models (Hopkins, 1982). A better treatment of the interdependence among units of observation is to employ an efficient statistical modeling technique known as hierarchical linear modeling (HLM). In a typical educational setting in which students are taught in classrooms, classrooms are embedded within schools, and schools are embedded within school districts. HLM can model data at three levels: students

at Level 1, classrooms at Level 2, and schools at Level 3. Each higher-level (say, Level 2) model takes into account the nested nature of data collected at a lower level (say, Level 1). Thus, the hierarchical nature of the data is completely and efficiently captured in HLM (Raudenbush & Bryk, 1988).

—Chao-Ying Joanne Peng

See also CROSS-SECTIONAL DESIGN

REFERENCES

- Hopkins, K. D. (1982). The unit of analysis: Group means versus individual observations. *American Educational Research Journal*, 19(1), 5–18.
- Huck, S. (2000). *Reading statistics and research* (3rd ed.). New York: Addison-Wesley Longman.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.
- Kirk, R. E. (1999). *Statistics: An introduction* (4th ed.). Orlando, FL: Harcourt Brace.
- Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.
- Raudenbush, S. W., & Bryk, A. S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. *Review of Research in Education*, 15, 423–476.

NETWORK ANALYSIS

Network analysis is the interdisciplinary study of social relations and has roots in anthropology, sociology, psychology, and applied mathematics. It conceives of social structure in relational terms, and its most fundamental construct is that of a *social network*, comprising at the most basic level a set of *social actors* and a set of *relational ties* connecting pairs of these actors. A primary assumption is that social actors are interdependent and that the relational ties among them have important consequences for each social actor as well as for the larger social groupings that they comprise.

The *nodes* or *members* of the network can be groups or organizations as well as people. Network analysis involves a combination of theorizing, model building, and empirical research, including (possibly) sophisticated data analysis. The goal is to study

network structure, often analyzed using such concepts as density, centrality, prestige, mutuality, and role. Social network data sets are occasionally multidimensional and/or longitudinal, and they often include information about *actor attributes*, such as actor age, gender, ethnicity, attitudes, and beliefs.

A basic premise of the social network paradigm is that knowledge about the structure of social relationships enriches explanations based on knowledge about the attributes of the actors alone. Whenever the social context of individual actors under study is relevant, relational information can be gathered and studied. Network analysis goes beyond measurements taken on individuals to analyze data on patterns of relational ties and to examine how the existence and functioning of such ties are constrained by the social networks in which individual actors are embedded. For example, one might measure the relations “communicate with,” “live near,” “feel hostility toward,” and “go to for social support” on a group of workers. Some network analyses are longitudinal, viewing changing social structure as an outcome of underlying processes. Others link individuals to events (affiliation networks), such as a set of individuals participating in a set of community activities.

Network structure can be studied at many different levels: the dyad, triad, subgroup, or even the entire network. Furthermore, network theories can be posited at a variety of different levels. Although this multilevel aspect of network analysis allows different structural questions to be posed and studied simultaneously, it usually requires the use of methods that go beyond the standard approach of treating each individual as an independent unit of analysis. This is especially true for studying a *complete* or whole network: a census of a well-defined population of social actors in which all ties, of various types, among all the actors are measured. Such analyses might study structural balance in small groups, transitive flows of information through indirect ties, structural equivalence in organizations, or patterns of relations in a set of organizations.

For example, network analysis allows a researcher to model the interdependencies of organization members. The paradigm provides concepts, theories, and methods to investigate how informal organizational structures intersect with formal bureaucratic structures in the unfolding flow of work-related actions of organizational members and in their evolving sets

of knowledge and beliefs. Hence, it has informed many of the topics of organizational behavior, such as leadership, attitudes, work roles, turnover, and computer-supported cooperative work.

HISTORICAL BACKGROUND

Network analysis has developed out of several research traditions, including (a) the birth of sociometry in the 1930s spawned by the work of the psychiatrist Jacob L. Moreno; (b) ethnographic efforts in the 1950s and 1960s to understand migrations from tribal villages to polyglot cities, especially the research of A. R. Radcliffe-Brown; (c) survey research since the 1950s to describe the nature of personal communities, social support, and social mobilization; and (d) archival analysis to understand the structure of interorganizational and international ties. Also noteworthy is the work of Claude Lévi-Strauss, who was the first to introduce formal notions of kinship, thereby leading to a mathematical algebraic theory of relations, and the work of Anatol Rapoport, perhaps the first to propose an elaborate statistical model of relational ties and flow through various nodes.

Highlights of the field include the adoption of sophisticated mathematical models, especially discrete mathematics and graph theory, in the 1940s and 1950s. Concepts such as transitivity, structural equivalence, the strength of weak ties, and centrality arose from network research by James A. Davis, Samuel Leinhardt, Paul Holland, Harrison White, Mark Granovetter, and Linton Freeman in the 1960s and 1970s. Despite the separateness of these many research beginnings, the field grew and was drawn together in the 1970s by formulations in graph theory and advances in computing. Network analysis, as a distinct research endeavor, was born in the early 1970s. Noteworthy in its birth are the pioneering text by Harary, Norman, and Cartwright (1965); the appearance in the late 1970s of network analysis software, much of it arising at the University of California, Irvine; and annual conferences of network analysts, now sponsored by the International Network for Social Network Analysis. These well-known "Sunbelt" Social Network Conferences now draw as many as 400 international participants. A number of fields, such as organizational science, have experienced rapid growth through the adoption of a network perspective.

Over the years, the social network analytic perspective has been used to gain increased understanding of many diverse phenomena in the social and behavioral sciences, including (taken from Wasserman & Faust, 1994)

- Occupational mobility
- Urbanization
- World political and economic systems
- Community elite decision making
- Social support
- Community psychology
- Group problem solving
- Diffusion and adoption of information
- Corporate interlocking
- Belief systems
- Social cognition
- Markets
- Sociology of science
- Exchange and power
- Consensus and social influence
- Coalition formation

In addition, it offers the potential to understand many contemporary issues, including

- The Internet
- Knowledge and distributed intelligence
- Computer-mediated communication
- Terrorism
- Metabolic systems
- Health, illness, and epidemiology, especially of HIV

Before a discussion of the details of various network research methods, we mention in passing a number of important measurement approaches.

MEASUREMENT

Complete Networks

In complete network studies, a census of network ties is taken for all members of a prespecified population of network members. A variety of methods may be used to observe the network ties (e.g., survey, archival, participant observation), and observations may be made on a number of different types of network tie. Studies of complete networks are often appropriate when it is desirable to understand the action of network members in terms of their location in a broader social

system (e.g., their centrality in the network, or more generally in terms of their patterns of connections to other network members). Likewise, it may be necessary to observe a complete network when properties of the network as a whole are of interest (e.g., its degree of centralization, fragmentation, or connectedness).

Ego-Centered Networks

The size and scope of complete networks generally preclude the study of all the ties and possibly all the nodes in a large, possibly unbounded population. To study such phenomena, researchers often use survey research to study a sample of personal networks (often called *ego-centered* or *local* networks). These smaller networks consist of the set of specified ties that links *focal persons* (or *egos*) at the centers of these networks to a set of close “associates” or *alters*. Such studies focus on an ego’s ties and on ties among ego’s alters. Ego-centered networks can include relations such as kinship, weak ties, frequent contact, and provision of emotional or instrumental aid. These relations can be characterized by their variety, content, strength, and structure. Thus, analysts might study network member *composition* (such as the percentage of women providing social or emotional support, for example, or basic actor attributes more generally); network characteristics (e.g., percentage of dyads that are mutual); measures of *relational association* (do strong ties with immediate kin also imply supportive relationships?); and network structure (how densely knit are various relations? do actors cluster in any meaningful way?).

SNOWBALL SAMPLING AND LINK TRACING STUDIES

Another possibility, to study large networks, is simply to sample nodes or ties. Sampling theory for networks contains a small number of important results (e.g., estimation of subgraphs or subcomponents; many originated with Ove Frank) as well as a number of unique techniques or strategies such as snowball sampling, in which a number of nodes are sampled, then those linked to this original sample are sampled, and so forth, in a multistage process. In a link-tracing sampling design, emphasis is on the links rather than the actors—a set of social links is followed from one respondent to another. For hard-to-access or hidden

populations, such designs are considered the most practical way to obtain a sample of nodes.

COGNITIVE SOCIAL STRUCTURES

Social network studies of *social cognition* investigate how individual network actors perceive the ties of others and the social structures in which they are contained. Such studies often involve the measurement of multiple perspectives on a network, for instance, by observing each network member’s view of who is tied to whom in the network. David Krackhardt referred to the resulting data arrays as *cognitive social structures*. Research has focused on clarifying the various ways in which social cognition may be related to network locations: (a) People’s positions in social structures may determine the specific information to which they are exposed, and hence, their perception; (b) structural position may be related to characteristic patterns of social interactions; (c) structural position may frame social cognitions by affecting people’s perceptions of their social locales.

METHODS

Social network analysts have developed methods and tools for the study of relational data. The techniques include graph theoretic methods developed by mathematicians (many of which involve counting various types of subgraphs); algebraic models popularized by mathematical sociologists and psychologists; and statistical models, which include the social relations model from social psychology and the recent family of random graphs first introduced into the network literature by Ove Frank and David Strauss. Software packages to fit these models are widely available.

Exciting recent developments in network methods have occurred in the statistical arena and reflect the increasing theoretical focus in the social and behavioral sciences on the interdependence of social actors in dynamic, network-based social settings. Therefore, a growing importance has been accorded the problem of constructing theoretically and empirically plausible parametric models for structural network phenomena and their changes over time. Substantial advances in statistical computing are now allowing researchers to more easily fit these more complex models to data.

SOME NOTATION

In the simplest case, network studies involve a single type of directed or nondirected tie measured for all pairs of a node set $N = \{1, 2, \dots, n\}$ of individual actors. The observed tie linking node i to node j ($i, j \in N$) can be denoted by x_{ij} and is often defined to take the value 1 if the tie is observed to be present and 0 otherwise. The network may be either *directed* (in which case x_{ij} and x_{ji} are distinguished and may take different values) or *nondirected* (in which case x_{ij} and x_{ji} are not distinguished and are necessarily equal in value). Other cases of interest include the following:

1. *Valued networks*, where x_{ij} takes values in the set $\{0, 1, \dots, C - 1\}$.
2. *Time-dependent networks*, where x_{ijt} represents the tie from node i to node j at time t .
3. *Multiple relational or multivariate networks*, where x_{ijk} represents the tie of type k from node i to node j (with $k \in R = \{1, 2, \dots, r\}$, a fixed set of *types* of tie).

In most of the statistical literature on network methods, the set N is regarded as fixed and the network ties are assumed to be random. In this case, the tie linking node i to node j may be denoted by the random variable X_{ij} and the $n \times n$ array $X = [X_{ij}]$ of random variables can be regarded as the adjacency matrix of a *random (directed) graph* on N . The state space of all possible realizations of these arrays is Ω_n . The array $x = [x_{ij}]$ denotes a realization of X .

GRAPH THEORETIC TECHNIQUES

Graph theory has played a critical role in the development of network analysis. Graph theoretical techniques underlie approaches to understanding cohesiveness, connectedness, and fragmentation in networks. Fundamental measures of a network include its *density* (the proportion of possible ties in the network that are actually observed) and the degree sequence of its nodes. In a nondirected network, the *degree* d_i of node i is the number of distinct nodes to which node i is connected ($d_i = \sum_{j \in N} x_{ij}$). In a directed network, sequences of *indegrees* ($\sum_{j \in N} x_{ji}$) and *outdegrees* ($\sum_{j \in N} x_{ij}$) are of interest. Methods for characterizing and identifying cohesive subsets in a network have depended on the notion of a *clique* (a subgraph of network nodes, every pair of which is connected) as well as on a variety of

generalizations (including *k-clique*, *k-plex*, *k-core*, *LS-set*, and *k-connected subgraph*).

Our understanding of connectedness, connectivity, and centralization is also informed by the distribution of path lengths in a network. A *path* of length k from one node i to another node j is defined by a sequence $i = i_1, i_2, \dots, i_{k+1} = j$ of distinct nodes such that i_h and i_{h+1} are connected by a network tie. If there is no path from i to j of length $n - 1$ or less, then j is not reachable from i and the distance from i to j is said to be infinite; otherwise, the distance from i to j is the length of the shortest path from i to j . A directed network is *strongly connected* if each node is reachable from each other node; it is *weakly connected* if, for every pair of nodes, at least one of the pair is reachable from the other. For nondirected networks, a network is connected if each node is reachable from each other node, and the *connectivity*, κ , is the least number of nodes whose removal results in a disconnected (or trivial) subgraph.

Graphs that contain many cohesive subsets as well as short paths, on average, are often termed *small world* networks, following early work by Stanley Milgram, and more recent work by Duncan Watts. Characterizations of the centrality of each actor in the network are typically based on the actor's degree (*degree centrality*), on the lengths of paths from the actor to all other actors (*closeness centrality*), or on the extent to which the shortest paths between other actors pass through the given actor (*betweenness centrality*). Measures of network *centralization* signify the extent of heterogeneity among actors in these different forms of centrality.

ALGEBRAIC TECHNIQUES

Closely related to graph theoretic approaches is a collection of algebraic techniques that has been developed to understand social roles and structural regularities in networks. Characterizations of role have developed in terms of mappings on networks, and descriptions of structural regularities have been facilitated by the construction of algebras among labeled network walks. An important proposition about what it means for two actors to have the same social role is embedded in the notion of *structural equivalence*: Two actors are said to be *structurally equivalent* if they relate to and are related by every other network actor in exactly the same way (thus, nodes i and j are

structurally equivalent if, for all $k \in N$, $x_{ik} = x_{jk}$ and $x_{ki} = x_{kj}$). Generalizations to automorphic and regular equivalence are based on more general mappings on N and capture the notion that similarly positioned network nodes are related to *similar* others in the same way.

Approaches to describing structural regularities in multiple networks have grown out of earlier characterizations of structure in kinship systems, and can be defined in terms of labeled walks in multiple networks. Two nodes i and j are connected by a labeled *walk* of type $k_1 k_2 \dots k_h$ if there is a sequence of nodes $i = i_1, i_2, \dots, i_{h+1} = j$ such that i_q is connected to i_{q+1} by a tie of type k_q (note that the nodes in the sequence need not all be distinct, so that a walk is a more general construction than a path). Each sequence $k_1 k_2 \dots k_h$ of tie labels defines a derived network whose ties signify the presence of labeled walks of that specified type among pairs of network nodes. Equality and ordering relations among these derived networks lead to various algebraic structures (including *semigroups* and partially ordered semigroups) and describe observed regularities in the structure of walks and paths in the multiple network. For example, *transitivity* in a directed network with ties of type k is a form of structural regularity associated with the observation that walks of type kk link two nodes only if the nodes are also linked by a walk of type k .

STATISTICAL TECHNIQUES

A simple statistical model for a (directed) graph assumes a BERNOULLI distribution, in which each edge, or tie, is statistically independent of all others and governed by a theoretical probability P_{ij} . In addition to edge independence, simplified versions also assume equal probabilities across ties; other versions allow the probabilities to depend on structural parameters. These distributions often have been used as models for at least 40 years, but are of questionable utility because of the independence assumption.

Dyadic Structure in Networks

Statistical models for social network phenomena have been developed from their edge-independent beginnings in a number of major ways. The p_1 model recognized the theoretical and empirical importance

of dyadic structure in social networks, that is, of the interdependence of the variables x_{ij} and x_{ji} . This class of Bernoulli dyad distributions and their generalization to valued, multivariate, and time-dependent forms gave parametric expression to ideas of reciprocity and exchange in dyads and their development over time. The model assumes that each dyad (x_{ij}, x_{ji}) is independent of every other and, in a commonly constrained form, specifies

$$P(X = x) = \pi_{i < j} \exp \left[\lambda_{ij} + \theta \left(\sum_{i < j} x_{ij} \right) + \rho \left(\sum_{i < j} x_{ij} x_{ji} \right) + \alpha_i \left(\sum_j x_{ij} \right) + \beta_j \left(\sum_i x_{ij} x_{ji} \right) \right], \quad (1)$$

where θ is a density parameter, ρ is a reciprocity parameter, the parameters α_i and β_j reflect individual differences in expansiveness and popularity, and λ_{ij} ensures that probabilities for each dyad sum to 1. This is a LOG-LINEAR MODEL and easily fit. Generalizations of this model are numerous, and include *stochastic block models*, representing hypotheses about the interdependence of social positions and the patterning of network ties; mixed models, such as p_2 ; and *latent space models* for networks.

Null Models for Networks

The assumption of dyadic independence is questionable. Thus, another series of developments has been motivated by the problem of assessing the degree and nature of departures from simple structural assumptions like dyadic independence. A number of *conditional uniform* random graph distributions were introduced as null models for exploring the structural features of social networks. These distributions, denoted by $U|Q$, are defined over subsets Q of the state space Ω_n of directed graphs and assign equal probability to each member of Q . The subset Q is usually chosen to have some specified set of properties (e.g., a fixed number of mutual, asymmetric, and null dyads). When Q is equal to Ω_n , the distribution is referred to as the *uniform* (di)graph distribution, and is equivalent to a Bernoulli distribution with homogeneous tie probabilities. Enumeration of the

members of Q and simulation of $U|Q$ are often straightforward, although certain cases, such as the distribution that is conditional on the indegree and outdegree of each node i in the network, require more complicated approaches.

A typical application of these distributions is to assess whether the occurrence of certain higher-order (e.g., triadic) features in an observed network is unusual, given the assumption that the data arose from a uniform distribution that is conditional on plausible lower-order (e.g., dyadic) features. This general approach has also been developed for the analysis of multiple networks. The best known example is probably Frank Baker and Larry Hubert's Quadratic Assignment Procedure (QAP) for networks. In this case, the association between two graphs defined on the same set of nodes is assessed using a uniform multi-graph distribution that is conditional on the unlabeled graph structure of each constituent graph.

Extradynamic Local Structure in Networks

A significant step in the development of parametric statistical models for social networks was taken by Frank and Strauss (1986) with the introduction of the class of *Markov random graphs*. This class of models permitted the parameterization of extradynamic local structural forms and so allowed a more explicit link between some important theoretical propositions and statistical network models. These models are based on the fact that the Hammersley-Clifford theorem provides a general probability distribution for X from a specification of which pairs (X_{ij}, X_{kl}) of tie random variables are conditionally dependent, given the values of all other random variables.

Specifically, define a *dependence graph* $\mathbf{D}$ with node set $N(\mathbf{D}) = \{(X_{ij}: i, j \in N, i \neq j)\}$ and edge set $E(\mathbf{D}) = \{(X_{ij}, X_{kl}): X_{ij} \text{ and } X_{kl} \text{ are assumed to be conditionally dependent, given the rest of } X\}$. Frank and Strauss used $\mathbf{D}$ to obtain a model for $\Pr(X = \mathbf{x})$, denoted p^* by later researchers, in terms of parameters and substructures corresponding to cliques of $\mathbf{D}$. The model has the form

$$\Pr(X = \mathbf{x}) = p^*(\mathbf{x}) = (1/c) \exp \left[\sum_{P \subseteq N(\mathbf{D})} \alpha_P z_P(\mathbf{x}) \right], \quad (2)$$

where

1. The summation is over all cliques P of $\mathbf{D}$ [with a *clique* of $\mathbf{D}$ defined as a nonempty subset P of $N(\mathbf{D})$ such that $|P| = 1$ or $(X_{ij}, X_{kl}) \in E(\mathbf{D})$ for all $X_{ij}, X_{kl} \in P$].
2. $z_P(\mathbf{x}) = \prod_{X_{ij} \in P} x_{ij}$ is the (observed) *network statistic* corresponding to the clique P of $\mathbf{D}$
3. $c = \sum_{\mathbf{x}} \exp \left\{ \sum_P \alpha_P z_P(\mathbf{x}) \right\}$ is a *normalizing* quantity.

One possible dependence assumption is Markov, in which $(X_{ij}, X_{kl}) \in E(\mathbf{D})$ whenever $\{i, j\} \cap \{k, l\} \neq \emptyset$. This assumption implies that the occurrence of a network tie from one node to another is conditionally dependent on the presence or absence of other ties in a *local neighborhood* of the tie. A Markovian local neighborhood for X_{ij} comprises all possible ties involving i and/or j . Many other dependence assumptions are also possible, and the task of identifying appropriate dependence assumptions in any modeling venture poses a significant theoretical challenge.

These random graph models permit the parameterization of many important ideas about local structure in univariate social networks, including transitivity, local clustering, degree variability, and centralization. Valued, multiple, and temporal generalizations also lead to parameterizations of substantively interesting multirelational concepts, such as those associated with balance and clusterability, generalized transitivity and exchange, and the strength of weak ties. Pseudo-maximum likelihood estimation is easy; maximum likelihood estimation is difficult, but not impossible.

Dynamic Models

A significant challenge is to develop models for the emergence of network phenomena, including the evolution of networks and the unfolding of individual actions (e.g., voting, attitude change, decision making) and interpersonal transactions (e.g., patterns of communication or interpersonal exchange) in the context of long-standing relational ties. Early attempts to model the evolution of networks in either discrete or continuous time assumed dyad independence and Markov processes in time. A step toward continuous time MARKOV CHAIN models for network evolution that relaxes the assumption of dyad independence has been taken by Tom Snijders and colleagues. This approach also illustrates the potentially

valuable role of simulation techniques for models that make empirically plausible assumptions; clearly, such methods provide a promising focus for future development. Computational models based on simulations are becoming increasingly popular in network analysis; however, the development of associated model evaluation approaches poses a significant challenge.

Current research (as of 2003), including future challenges, such as statistical estimation of complex model parameters, model evaluation, and dynamic statistical models for longitudinal data, can be found in Carrington, Scott, and Wasserman (2003). Applications of the techniques and definitions mentioned here can be found in Scott (1992) and Wasserman and Faust (1994).

—Stanley Wasserman and Philippa Pattison

AUTHOR'S NOTE: Research supported by the U.S. Office of Naval Research and the Australian Research Council.

REFERENCES

- Boyd, J. P. (1990). *Social semigroups: A unified theory of scaling and blockmodeling as applied to social networks*. Fairfax, VA: George Mason University Press.
- Carrington, P. J., Scott, J., & Wasserman, S. (Eds.). (2003). *Models and methods in social network analysis*. New York: Cambridge University Press.
- Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81, 832–842.
- Friedkin, N. (1998). *A structural theory of social influence*. New York: Cambridge University Press.
- Harary, F., Norman, D., & Cartwright, D. (1965). *Structural models for directed graphs*. New York: Free Press.
- Monge, P., & Contractor, N. (2003). *Theories of communication networks*. New York: Oxford University Press.
- Pattison, P. E. (1993). *Algebraic models for social networks*. New York: Cambridge University Press.
- Scott, J. (1992). *Social network analysis*. London: Sage.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.
- Wasserman, S., & Galaskiewicz, J. (Eds.). (1994). *Advances in social network analysis*. Thousand Oaks, CA: Sage.
- Watts, D. (1999). *Small worlds: The dynamics of networks between order and randomness*. Princeton, NJ: Princeton University Press.
- Wellman, B., & Berkowitz, S. D. (Eds.). (1997). *Social structures: A network approach* (updated ed.). Greenwich, CT: JAI.

NEURAL NETWORK

Neural networks are adaptive statistical models based on an analogy with the structure of the brain. They are *adaptive* because they can learn to estimate the PARAMETERS of some population using a small number of exemplars (one or a few) at a time. They do not differ *essentially* from standard statistical MODELS. For example, one can find neural network architectures akin to DISCRIMINANT ANALYSIS, PRINCIPAL COMPONENTS ANALYSIS, LOGISTIC REGRESSION, and other techniques. In fact, the same mathematical tools can be used to analyze standard statistical models and neural networks. Neural networks are used as *statistical tools* in a variety of fields, including psychology, statistics, engineering, econometrics, and even physics. They are used also as *models* of cognitive processes by neuro- and cognitive scientists.

Basically, neural networks are built from simple units, sometimes called *neurons* or cells by analogy with the real thing. These units are linked by a set of weighted connections. Learning is usually accomplished by modification of the connection weights. Each unit CODES or corresponds to a feature or a characteristic of a pattern that we want to analyze or that we want to use as a PREDICTOR VARIABLE.

These networks usually organize their units into several layers. The first layer is called the *input* layer, and the last one is the *output* layer. The intermediate layers (if any) are called the *hidden* layers. The information to be analyzed is fed to the neurons of the first layer and then propagated to the neurons of the second layer for further processing. The result of this processing is then propagated to the next layer and so on until the last layer. Each unit receives some information from other units (or from the external world through some devices) and processes this information, which will be converted into the output of the unit.

The goal of the network is to learn or to discover some association between input and output patterns, or to analyze, or to find the structure of the input patterns. The learning process is achieved through the modification of the connection weights between units. In statistical terms, this is equivalent to interpreting the value of the connections between units as parameters (e.g., like the values of a and b in the REGRESSION equation $y = a + bx$) to be estimated. The learning

process specifies the “ALGORITHM” used to estimate the parameters.

THE BUILDING BLOCKS OF NEURAL NETWORKS

Neural networks are made of basic units (see Figure 1) arranged in layers. A unit collects information provided by other units (or by the external world) to which it is connected with *weighted* connections called *synapses*. These weights, called *synaptic weights*, multiply (i.e., amplify or attenuate) the input information: A positive weight is considered excitatory, a negative weight inhibitory.

Each of these units is a simplified model of a neuron and transforms its input information into an output response. This transformation involves two steps: First, the activation of the neuron is computed as the weighted sum of its inputs, and second, this activation is transformed into a response by using a *transfer* function. Formally, if each input is denoted x_i , and each weight w_i , then the activation is equal to $a = \sum x_i w_i$, and the output, denoted o , is obtained as $o = f(a)$. Any function whose domain is real numbers can be used as a transfer function. The most popular ones are the linear function ($o \propto a$); the step function (activation values less than a given threshold are set to 0 or to -1 and the other values are set to $+1$); the logistic function $[f(x) = \frac{1}{1+\exp\{-x\}}]$, which maps the real numbers into the interval $[-1 + 1]$ and whose derivative, needed for learning, is easily computed $\{f'(x) = f(x)[1 - f(x)]\}$; and the normal or Gaussian function $[o = (\sigma\sqrt{2\pi})^{-1} \times \exp\{-\frac{1}{2}(a/\sigma)^2\}]$. Some of these functions can include

probabilistic variations; for example, a neuron can transform its activation into the response $+1$ with a probability of $\frac{1}{2}$ when the activation is larger than a given threshold.

The architecture (i.e., the pattern of connectivity) of the network, along with the transfer functions used by the neurons and the synaptic weights, completely specify the behavior of the network.

LEARNING RULES

Neural networks are adaptive statistical devices. This means that they can change iteratively the values of their parameters (i.e., the synaptic weights) as a function of their performance. These changes are made according to *learning rules*, which can be characterized as *supervised* (when a desired output is known and used to compute an error signal) or *unsupervised* (when no such error signal is used).

The Widrow-Hoff (also called the gradient descent or delta rule) is the most widely known supervised learning rule. It uses the difference between the actual input of the cell and the desired output as an error signal for units in the output layer. Units in the hidden layers cannot compute directly their error signal, but they estimate it as a function (e.g., a weighted average) of the error of the units in the following layer. This adaptation of the Widrow-Hoff learning rule is known as *error backpropagation*. With Widrow-Hoff learning, the correction to the synaptic weights is proportional to the error signal multiplied by the value of the activation given by the derivative of the transfer function. Using the derivative has the effect of making finely tuned corrections when the activation is near its extreme values (minimum or maximum) and larger corrections when the activation is in its middle range. Each correction has the immediate effect of making the error signal *smaller* if a similar input is applied to the unit. In general, supervised learning rules implement optimization algorithms akin to *descent* techniques because they search for a set of values for the free parameters (i.e., the synaptic weights) of the system such that some error function computed for the whole network is minimized.

The Hebbian rule is the most widely known unsupervised learning rule. It is based on work by the Canadian neuropsychologist Donald Hebb, who theorized that neuronal learning (i.e., synaptic change) is a local phenomenon expressible in terms of the temporal correlation between the activation values of neurons. The

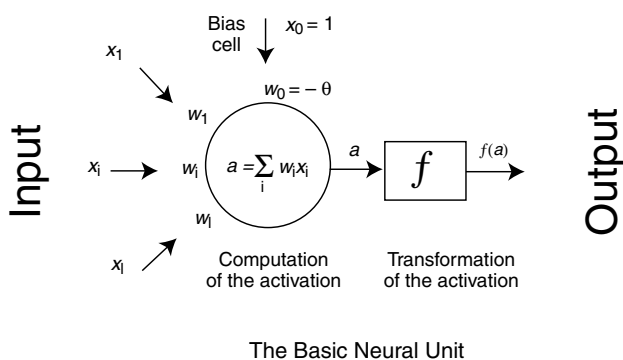


Figure 1 The Basic Neural Unit Processes the Input Information Into the Output Information

synaptic change depends on both the presynaptic and activities and states that the change in a synaptic weight is a function of the temporal correlation between the presynaptic and postsynaptic activities. Specifically, the value of the synaptic weight between two neurons increases whenever they are in the same state and decreases when they are in different states.

SOME IMPORTANT NEURAL NETWORK ARCHITECTURE

One the most popular architectures in neural networks is the *multilayer perceptron* (see Figure 2). Most of the networks with this architecture use the Widrow-Hoff rule as their learning algorithm and the logistic function as the transfer function of the units of the hidden layer (the transfer function is, in general, nonlinear for these neurons). These networks are very popular because they can approximate any multivariate function relating the input to the output. In a statistical framework, these networks are akin to MULTIVARIATE NONLINEAR REGRESSION. When the input patterns are the same as the output patterns, these networks are called *auto-associators*. They are closely related to linear (if the hidden units are linear) or nonlinear (if not) PRINCIPAL COMPONENT ANALYSIS and other statistical techniques linked to the GENERAL LINEAR MODEL (see Abdi, Valentin, Edelman, & O'Toole, 1996), such as DISCRIMINANT ANALYSIS or CORRESPONDENCE ANALYSIS.

A recent development generalizes the *radial basis function* (RBF) networks (see Abdi, Valentin, & Edelman, 1999) and integrates them with *statistical learning theory* (see Vapnik, 1999) under the name of *support vector machine* or SVM (see Schölkopf & Smola, 2003). In these networks, the hidden units (called the support vectors) represent possible (or even real) input patterns, and their response is a function of their similarity to the input pattern under consideration. The similarity is evaluated by a *kernel* function (e.g.,

dot product; in the radial basis function, the kernel is the Gaussian transformation of the Euclidean distance between the support vector and the input). In the specific case of RBF networks—which we will use as an example of SVM—the output of the units of the hidden layers are connected to an output layer composed of linear units. In fact, these networks work by breaking the difficult problem of a nonlinear approximation into two more simple ones. The first step is a simple nonlinear mapping (the Gaussian transformation of the distance from the kernel to the input pattern), and the second step corresponds to a linear transformation from the hidden layer to the output layer. Learning occurs at the level of the output layer. The main difficulty with these architectures resides in the choice of the support vectors and the specific kernels to use. These networks are used for pattern recognition, classification, and clustering data.

VALIDATION

From a statistical point of view, neural networks represent a class of nonparametric adaptive models. In this framework, an important issue is to evaluate the performance of the model. This is done by separating the data into two sets: the training set and the testing set. The parameters (i.e., the value of the synaptic weights) of the network are computed using the training set. Then, learning is stopped and the network is evaluated with the data from the testing set. This cross-validation approach is akin to BOOTSTRAPPING or the JACKKNIFE METHOD.

USEFUL REFERENCES

Neural network theory connects several domains from the neurosciences to engineering, including statistical theory. This diversity of sources also creates a real heterogeneity in the presentation of the material, because textbooks often try to address only one portion of the interested readership. The following references should be helpful to the reader interested in the statistical properties of neural networks: Abdi et al. (1999), Bishop (1995), Cherkassky and Mulier (1998), Duda, Hart, and Stork (2001), Hastie, Tibshirani, and Friedman (2001), Looney (1997), Ripley (1996), and Vapnik (1999).

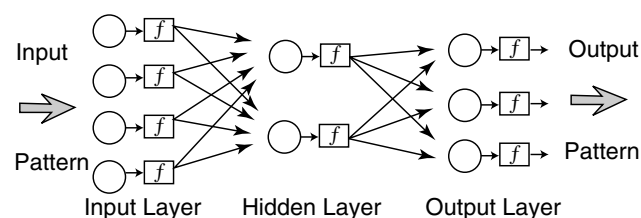


Figure 2 A Multilayer Perceptron

—Hervé Abdi

REFERENCES

- Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks*. Thousand Oaks, CA: Sage.
- Abdi, H., Valentin, D., Edelman, B., & O'Toole, A. J. (1996). A Widrow-Hoff learning rule for a generalization of the linear auto-associator. *Journal of Mathematical Psychology*, 40, 175–182.
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.
- Cherkassky, V., & Mulier, F. (1998). *Learning from data*. New York: Wiley.
- Duda, R., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Wiley.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer-Verlag.
- Looney, C. G. (1997). *Pattern recognition using neural networks*. Oxford, UK: Oxford University Press.
- Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge, UK: Cambridge University Press.
- Schölkopf, B., & Smola, A. J. (2003). *Learning with kernel*. Cambridge: MIT Press.
- Vapnik, V. N. (1999). *Statistical learning theory*. New York: Wiley.

NOMINAL VARIABLE

This variable is simply a categorization of cases. It allows us to show that the cases are different in terms of the categories, and no more than that. For example, unlike an ordinal variable, we are not able to rank order the categories in any way. An example of a nominal variable is the voting behavior of a sample. We can classify members of the sample in terms of the political party for which they voted at the most recent election, but we cannot infer any more about the differences between people in terms of that classification.

—Alan Bryman

See also ATTRIBUTE, CATEGORICAL, DISCRETE

NOMOTHETIC. See NOMOTHETIC/IDEOGRAPHIC

NOMOTHETIC/IDEOGRAPHIC

The distinction between nomothetic, which pertains to the construction of general models and laws

in science, and ideographic, which is concerned with the detailed understanding of particular circumstances, was proposed by Wilhelm Windleband in 1894 (Smith, 1998, pp. 141–142). Although intended to demarcate the natural (nomothetic) from the cultural (ideographic) sciences, the positioning of the demarcation, and even whether one can talk of such, has been controversial ever since.

Windleband held that although all science and knowledge referred to the same reality, different disciplines will have distinct concerns and view reality accordingly. Although this was not an assertion of a divide between “mind” and “matter” (see NATURALISM), he nevertheless thought that, in practice, there is a divide between the cultural world—produced by autonomous, self-reflecting human agents—and the physical world. The latter may be known objectively by human observers, but the social world must be studied subjectively through a strategy of interpretation.

The debates of the mid to late 20th century between INTERPRETIVISTS and POSITIVISTS have their origins in Windleband's distinction, although ironically, one of the best arguments for a fluidity of any divide between the nomothetic and ideographic comes from one of the key figures in interpretivist social science, Max Weber (1975). He saw the distinction not as a scientific versus nonscientific mode of inquiry; rather, he saw both as forms of scientific inquiry. The former Weber equates with abstract generalizations, as in law-like statements. The second he regarded as a science of concrete reality, of specific instances.

If one takes Weber's view, then virtually all of the sciences, natural and social, have both nomothetic and ideographic characteristics, and their forms of inquiry reflect this. Astronomers, for example, will provide detailed individual descriptions of particular bodies, but explain them in the context of abstract generalizations about wider phenomena. Likewise, an interpretation of a specific human interaction will require a contextualization within a wider social setting. The latter requires at least some form of taxonomy and idealization that can be grounded in ideographic interpretation and description. A statistical claim such as “20% of pensioners lived in poverty” can tell us nothing about any given individual pensioner, only pensioners in the abstract, yet the description “poverty” is likely to have been derived from an examination of specific, individual cases of poverty.

The maintenance of a clear divide has been challenged from another direction, that of the philosophy

and sociology of natural science. Windleband's original distinction, and, to an extent, Weber's redefinition, left untouched the positivistic view of natural science as value free and phenomenalist. However, since the Popper-Kuhn "revolution" of the early 1960s (see Lakatos & Musgrave, 1965), most accept that, at least to some degree, subjectivity must enter natural science through moral positioning, theory choice, and observation. Finally, in recent years, complex approaches to understanding physical and social systems (see Byrne, 1998) have emphasized the ontological continuities and similarities between such systems and consequently have raised objections to the epistemological and methodological divide of the kind proposed by Windleband.

—Malcolm Williams

REFERENCES

- Byrne, D. (1998). *Complexity theory and the social sciences: An introduction*. London: Routledge.
- Lakatos, I., & Musgrave, A. (Eds.). (1970). *Criticism and the growth of knowledge*. Cambridge, UK: Cambridge University Press.
- Smith, M. (1998). *Social science in question*. London: Sage.
- Weber, M. (1975). *Roscher and Knies*. New York: Free Press.

NONADDITIVE

Nonadditive is a term used to describe a function that relates one set of scores (e.g., scores on a variable, Y) to two or more other sets of scores (e.g., scores on variables X_1 and X_2). The relationship between Y and the other variables is said to be additive if Y can be expressed as a sum of those other variables. The sum can be a weighted sum and has the general expression

$$Y = a + w_1X_1 + w_2X_2 + \cdots + w_kX_k,$$

where a and the various w s are constants.

In some applications, restrictions are placed on the constants. For example, if $a = 0$ and each weight is set to 1.0, then Y is literally the sum of the X variables. Another type of restriction that a researcher might apply is that the weights must sum to 1.0 in accordance with the formulation

$$w_j = c_j / \sum c_j,$$

where c_j is an absolute weight for variable X_j . The summation is across the k weights. As an example,

suppose that Y is an additive function of X_1 and X_2 , and that $a = 0$, $c_1 = 1$, and $c_2 = 1$. Then, $w_1 = 1/(1 + 1) = 0.50$, and $w_2 = 1/(1 + 1) = 0.50$, yielding

$$Y = 0 + 0.50X_1 + 0.50X_2.$$

Note that this expression is equivalent to

$$Y = (X_1 + X_2)/2,$$

so that Y is the average of X_1 and X_2 . This example shows that expressing Y as an average of a set of variables is a type of additive function.

A nonadditive function is one where Y cannot be expressed as a (weighted) sum of the other variables. Nonadditive functions can take many forms—indeed, there is an infinite number of such functions. An example of a nonadditive function is a MULTIPLICATIVE function, where Y is said to equal the product of two variables, X_1 and X_2 .

Traditional MULTIPLE REGRESSION ANALYSIS is based on an additive model,

$$Y = \alpha + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_kX_k,$$

where α is the intercept and the various β s are the REGRESSION COEFFICIENTS.

—James J. Jaccard

REFERENCE

- Anderson, N. H. (1981). *Methods of information integration theory*. New York: Academic Press.

NONLINEAR DYNAMICS

Nonlinear dynamics refers to a broad range of behavior that can occur with many varieties of mathematical models that are structured with respect to time. Nonlinear dynamics can occur with both linear and nonlinear algebraic specifications, as with $dy/dt = ay$ or $dy/dt = ay^2$, respectively, where the parameter " a " is a constant. Minimally, the basic idea is that change occurs over time with respect to one or more variables such that this change cannot be represented as a straight line on a graph that places time on the horizontal axis. More specifically, nonlinear variation is change that does not follow an incremental pattern in which the same increment of change recurs with

every sequential and equally spaced change in time. The most frequent uses of the term *nonlinear dynamics* refer to more complex situations in which phenomena with longitudinal change are modeled using nonlinear algebraic formulations (often involving systems of interdependent equations with more than one variable) that either implicitly or explicitly reference time. A thorough mathematical introduction to nonlinear dynamics as it appears in both linear and nonlinear mathematical specifications can be found in Hirsch and Smale (1974). Treatments and examples of this same subject matter from a social science perspective can be found in Brown (1995a, 1995b, 1991).

In the physical and natural sciences, nonlinear dynamics is normally encountered using continuous-time differential equation model specifications, which is a consequence of the continuous-time processes of change that are naturally encountered in these fields. Exceptions are not rare, however, especially in the biological sciences, in which generational or seasonal changes are the focus of study. In those cases, difference equation approaches are sometimes employed. In the social sciences, nonlinear dynamics is often modeled using discrete-time difference equation structures, almost always due to the manner in which social scientific data are collected (e.g., periodic elections, decade-spaced census reports, periodically spaced survey data, etc.). Exceptions do occur here as well, and the technical difficulties involved in using continuous-time models with discretely measured data sets as are commonly encountered in the social sciences now have clear solutions (e.g., see Brown, 1995b). The choice of using a continuous- or discrete-time approach to modeling social phenomena has substantive consequences that can be important in certain settings, and a discussion of these substantive consequences can be found in Brown (1995b, pp. 13–30).

Nonlinear dynamics is normally described in terms of the following processes of change: regular, periodic, chaotic, and catastrophe. A regular process begins with a bifurcation, or a point in which a new process of change begins that differs structurally from a previous process of change. This bifurcation is followed by growth that normally experiences positive feedback and that later changes to negative feedback as the growth process slows. The classic example of a regular process involving nonlinear dynamics is the logistic equation $dy/dt = ay(k - y)$, where growth in the variable y continues smoothly as its values asymptotically approach the equilibrium value k . In the

neighborhood of an equilibrium, the growth process eventually approaches zero net change. Because of other processes of change that are temporarily external to this regular process, the regular process eventually becomes “ripe” for experiencing another bifurcation, which can lead to the initiation of a new process of change that can be any of the four listed above (including a new regular process).

A periodic process is one in which the values of the relevant variables recur at specified intervals. Periodic processes normally have periodic limit cycles, as compared with fixed-point equilibria that are typically associated with regular processes. A chaotic process (see CHAOS THEORY) differs from a periodic process in that changes in the values of the relevant variables never recur exactly, and this lack of repetition is not a consequence of a stochastic element. Chaotic processes often exhibit dramatically diverse variations in variable values. Thus, chaotic processes lack periodic limit cycles, although variable variations typically “hover” within an identifiable neighborhood of an unstable equilibrium (a “strange attractor”).

Catastrophe processes (see CATASTROPHE THEORY) experience sudden and dramatic changes that depart from previously existing dynamic processes. An earthquake is an obvious candidate for a phenomenon that can be modeled as a catastrophe process because incremental (i.e., regular process) changes in the positions of tectonic plates eventually lead to a sudden departure from a previously established, pressure-defined equilibrium. In essence, a catastrophe process is such that the previously dominant equilibrium and its basin disappears, and a new equilibrium and its basin becomes dominant for the dynamical system. Examples of catastrophe models using social scientific data can be found in Brown (1995a, 1995b).

—Courtney Brown

REFERENCES

- Brown, C. (1995a). *Chaos and catastrophe theories*. Thousand Oaks, CA: Sage.
- Brown, C. (1995b). *Serpents in the sand: Essays on the nonlinear nature of politics and human destiny*. Ann Arbor: University of Michigan Press.
- Brown, C. (1991). *Ballots of tumult: A portrait of volatility in American voting*. Ann Arbor: University of Michigan Press.
- Hirsch, M. W., & Smale, S. (1974). *Differential equations, dynamical systems, and linear algebra*. New York: Academic Press.

NONLINEARITY

Nonlinearity refers to a **RELATIONSHIP** between two or more variables that is not linear. A nonlinear relationship *cannot* be represented by a line in two dimensions or by a plane in three dimensions. A linear relationship, for two variables Y and X , is defined by the equation for a line, $Y = a + bX$, where a and b are the intercept and slope coefficients, respectively. However, in the social sciences, relationships are rarely deterministic and usually are specified in a **REGRESSION** model that includes an error term to account for unexplained variation in Y . The equation for a two-variable **LINEAR REGRESSION** model, specifying a linear relationship between Y and X_1 with the error term ε , and the intercept and slope **PARAMETERS** β_0 and β_1 , is $Y = \beta_0 + \beta_1 X_1 + \varepsilon$. The general form of a **MULTIVARIATE** linear regression model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$, which specifies a linear relationship between each X and Y if other X s are held constant.

Nonlinearity in regression models is any other pattern of relationship *not* defined by these linear **MODELS**. Therefore, nonlinearity can take many forms. It may be curvilinear, represented by a quadratic polynomial, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \varepsilon$, or a cubic polynomial, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_3^3 + \varepsilon$. It may involve transcendental **FUNCTIONS** such as natural **LOGARITHMS** ($\ln$) (e.g., $Y = \beta_0 + \beta_1 X_1 + \beta_2 \ln X_2 + \varepsilon$), or exponentials (e) (e.g., $Y = \beta_0 e^{\beta_1 X_1} + \varepsilon$). It may be **MULTIPLICATIVE**, such as this **LOG-LINEAR MODEL**, $Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \cdots X_k^{\beta_k} e^{\varepsilon}$.

It is important to differentiate between nonlinear models that are “intrinsically linear” and those that are “intrinsically nonlinear.” Intrinsically linear

models are nonlinear models that can be reduced to linear models by means of **TRANSFORMATIONS** on Y and/or the X s. This amounts to simply substituting the transformed variables for the original variables into a linear regression model. Common intrinsically linear models and their linearizing transformations are reported in Table 1.

A nonlinear model that is intrinsically linear is nonlinear with respect to the variables but linear with respect to the **PARAMETERS** to be estimated. In contrast, intrinsically nonlinear models are nonlinear with respect to the variables and the parameters. Once transformed, the parameters of intrinsically linear models can be estimated using **ORDINARY LEAST SQUARES** (**OLS**). Intrinsically nonlinear models, on the other hand, require more elaborate techniques for estimating the parameters. Often, they do not have unique parameter estimates that minimize the **SUM OF SQUARED ERRORS** (**SSE**), as in **OLS**. Instead, the parameters are estimated iteratively through a series of improved guesses.

As demonstrated in Table 1, the nonlinear model, $Y = \beta_0 e^{\beta_1 X_1} \varepsilon$, can be linearized by taking natural logarithms of both sides, $\ln Y' = \ln \beta_0 + \beta_1 X_1 + \ln \varepsilon$. However, if the nonlinear model assumes an additive error term instead of a multiplicative error term, $Y = \beta_0 e^{\beta_1 X_1} + \varepsilon$, the model cannot be linearized and must be estimated with nonlinear techniques instead of **OLS**. Another common example of an intrinsically nonlinear model is the **LOGISTIC REGRESSION** model with a **BINARY** outcome variable,

$$Y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}} + \varepsilon.$$

—Jani S. Little

Table 1 Linearizing Transformations for Nonlinear Models

Nonlinear Form	Transformation(s) to Linearize	Linear Form
$Y = \beta_0 + \beta_1 X_1^2 + \varepsilon$	$X'_1 = X_1^2$	$Y = \beta_0 + \beta_1 X'_1 + \varepsilon$
$Y = \beta_0 + \beta_1 \ln X_1 + \varepsilon$	$X'_1 = \ln X_1$	$Y = \beta_0 + \beta_1 X'_1 + \varepsilon$
$Y = \beta_0 e^{\beta_1 X_1} \varepsilon$	$Y' = \ln Y$ $\beta'_0 = \ln \beta_0$ $\varepsilon' = \ln \varepsilon$	$Y' = \beta'_0 + \beta_1 X_1 + \varepsilon'$
$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} e^{\varepsilon}$	$Y' = \ln Y$ $\beta'_0 = \ln \beta_0$ $X'_1 = \ln X_1$ $X'_2 = \ln X_2$	$Y' = \beta'_0 + \beta_1 X'_1 + \beta_2 X'_2 + \varepsilon$

REFERENCES

- Devore, J. L. (1991). *Probability and statistics for engineering and the sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
- Greene, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Hamilton, L. C. (1992). *Regression with graphics: A second course in applied statistics*. Pacific Grove, CA: Brooks/Cole.
- Kmenta, J. (1986). *Elements of econometrics* (2nd ed.). New York: Macmillan.

NONPARAMETRIC RANDOM-EFFECTS MODEL

Random-effects modeling is one of the several alternative approaches to deal with dependent observations such as those that occur in repeated measures or multilevel data structures. Nonparametric random effects models differ from standard (parametric) random effects models in that no assumptions are made about the distribution of the random effects. Actually, this is a form of LATENT CLASS ANALYSIS: The mixing distribution is modeled by means of a finite mixture structure. Early references to the nonparametric approach are Laird (1978) and Heckman and Singer (1982).

Let y_{ij} denote the response variable of interest, where the index j refers to a group and i to a replication within a group. Note that with repeated measures, groups and replications refer to individuals and time points. It is easiest to explain the random effects models within a GENERALIZED LINEAR MODELING framework; that is, to assume that the response variable comes from a distribution belonging to the exponential family and that the expectation of y_{ij} , $E(y_{ij})$, is modeled via a linear function after an appropriate transformation $g[\cdot]$.

A simple random intercept model without predictors has the following form:

$$g[E(y_{ij})] = \alpha_j.$$

To utilize a parametric approach, a distributional form is specified for α_j , typically normal: $\alpha_j \sim N(\mu, \tau^2)$. The unknown parameters to be estimated are the mean, μ , and the variance, τ^2 . An equivalent parameterization is $g[E(y_{ij})] = \mu + \tau u_j$, with $u_j \sim N(0, 1)$.

A nonparametric approach characterizes the distribution of α_j by an unspecified discrete mixing

distribution with C nodes (latent classes), where a particular latent class (LC) is enumerated by x , $x = 1, 2, \dots, C$. The intercept associated with class x is denoted by α_x and the size of class x by $P(x)$. The α_x and $P(x)$ are sometimes referred to as the location and weight of node x . The nonparametric model can be specified as follows:

$$g[E(y_{ij}|x)] = \alpha_x.$$

The nonparametric maximum likelihood estimator is obtained by increasing the number of latent classes until a saturation point is reached. In practice, however, researchers prefer solutions with less than the maximum number of classes.

The similarity between the parametric and nonparametric approaches becomes clear when one realizes that the α_x and $P(x)$ parameters can be used to compute the mean (μ) and the variance (τ^2) of the random effects, which are the unknown parameters in the parametric approach. Using elementary statistics, we get $\mu = \sum_{x=1}^C \alpha_x P(x)$ and $\tau^2 = \sum_{x=1}^C (\alpha_x - \mu)^2 P(x)$.

A more general model is obtained by including predictors z_{ijk} , yielding a two-level regression model with a random intercept:

$$g[E(y_{ij}|\mathbf{z}_{ij}, x)] = \alpha_x + \sum_{k=1}^K \beta_k z_{ijk}.$$

A special case is the (semiparametric) Rasch model, which is obtained by adding a dummy predictor for each replication i .

Also, the regression coefficient can be allowed to differ across latent classes, which is analogous to having random slopes in a MULTILEVEL ANALYSIS. This yields

$$g[E(y_{ij}|\mathbf{z}_{ij}, x)] = \alpha_x + \sum_{k=1}^K \beta_{kx} z_{ijk},$$

a model that is often referred to as a latent class (LC) or mixture regression model. In fact, this is one of the most important applications of latent class analysis: Unobserved subgroups are identified that differ with respect to the parameters of the regression model of interest.

Two computer programs for estimating LC regression models are GLIMMIX and Latent GOLD.

—Jeroen K. Vermunt and Jay Magidson

REFERENCES

- Heckman, J. J., & Singer, B. (1982). Population heterogeneity in demographic models. In K. Land & A. Rogers (Eds.), *Multidimensional mathematical demography*. New York: Academic Press.
- Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixture distribution. *Journal of the American Statistical Association*, 73, 805–811.
- Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86, 96–107.
- Wedel, M., & DeSarbo, W. S. (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 352–388). Cambridge, UK: Basil Blackwell.

NONPARAMETRIC REGRESSION.

See LOCAL REGRESSION

NONPARAMETRIC STATISTICS

Experimental, QUASI-EXPERIMENTAL, ex post facto, and correlation designs often generate results that are not continuous data (INTERVAL or ratio), but nominal or ordinal data instead. Thus, parametric statistics are not the appropriate tools to resolve the hypotheses or describe associations between variables. There is a need for tests and measures that basically do not have all the constraints of parametric ones, some for which there are no demands for continuous data, or NORMAL DISTRIBUTIONS for this data (thus, they are sometimes described as “distribution-free”). Also, there is a need for tests where group characteristics are nominal, the data consisting of the frequency of subjects having specified traits. Not surprisingly, these are referred to as nonparametric tests, and they are appropriate for analyzing nominal and rank/ordinal data in order to make inferences about POPULATIONS based upon REPRESENTATIVE SAMPLES.

NATURE OF DATA AND SOURCES

Nonparametric tests do have some requirements of their own, and their disadvantage is that their use is more likely to incur a TYPE II ERROR than a

Table 1 Criteria for Choosing Between Parametric and Nonparametric Tests

<i>Parametric</i>	<i>Nonparametric</i>
Continuous data (scores), and	Nominal data (frequencies in categories), or
Normal distributions for all groups	Ordinal data (ranks), or Non-normal distributions of continuous data

parallel parametric test. They are often rated against comparable parametric tests in terms of *power efficiency*. Siegel and Castellan (1988) describe a test that has a power efficiency of 90%: “When all the conditions of the parametric statistical test are satisfied the appropriate parametric test would be just as effective with a sample which is 10% smaller than that used in the non-parametric analysis” (p. 36). Thus, one could conceivably replace parametric tests with nonparametric counterparts by degrading the data from continuous to ranks, but at a cost to the power of the test. Table 1 summarizes the requirements for the two types of statistics.

Having noted some of the differences between parametric and nonparametric tests and the limitations of the latter, it is worth noting the advantages of using them. To summarize Siegel and Castellan (1988), the advantages are that such tests

1. Are applicable to small samples where the underlying distribution is not known.
2. May be more appropriate for the research question.
3. Contend with ranked data.
4. Contend with categorical (nominal) data.

TESTS OF SIGNIFICANT DIFFERENCE

Research designs often wish to determine whether the differences between two groups (or three or more groups) is greater than what could be attributed to natural variation. In other words, no matter how random a set of samples is from a population, the samples will not all be the same. When they are too different, they are not considered part of the same population. Although this is reasonably easy to understand when comparing means for scores for groups of subjects using parametric tests, it can be more difficult to comprehend with nonparametric counterparts, and they do not follow a

Table 2 Comparing Some Typical Parametric and Nonparametric Tests of Significance

<i>Independent Variables or Groups</i>					
<i>Measure of Dependent Variable</i>	<i>Two Groups</i>			<i>Three Groups</i>	
	<i>One-Sample</i>	<i>Independent</i>	<i>Related/ Matched</i>	<i>Independent</i>	<i>Related/ Matched</i>
Interval/Ratio	<i>z</i> -test	<i>t</i> -tests	<i>t</i> _{related} -test	One-way ANOVA	Randomized block ANOVA
Ordinal	<i>t</i> -test (<i>n</i> < 30)	Mann-Whitney U-test	Wilcoxon signed ranks test	Factorial ANOVA	ANCOVA
	Kolmogorov- Smirnov test			Kruskal-Wallis one-way analysis of variance	Friedman two-way analysis of variance by ranks
Nominal	χ^2 -test, goodness-of-fit	χ^2 -test, <i>k</i> × 2 tables	McNemar change test	χ^2 -test for <i>m</i> × <i>k</i> tables	Cochran Q-test

Table 3 Contrasting Some Typical Parametric and Nonparametric Measures of Correlation and Association

			<i>Nominal</i>	
	<i>Interval/Ratio</i>	<i>Ordinal</i>	<i>m</i> ≥ 2	<i>m</i> = 2
Interval/Ratio	Pearson product moment, r_{xy} (linear relationship). If nonlinear, monotonic: both reduced to ordinal	(Interval/ratio variable reduced to ordinal)		Point biserial, r_{pb}
Ordinal		↓ Spearman's rho, r_s (linear relationship)	(Ordinal variable reduced to nominal)	
Nominal			↓ Cramer's C (V, ϕ_c)	
<i>k</i> ≥ 2				
<i>k</i> = 2				Phi, ϕ (dichotomies)

common theme. The differences of interest for these tests typically can be differences in

- *Proportions* or frequencies in categories of samples as compared to what would be expected (e.g., is the proportion of male to female nurses consistent with the national pattern?)
- *Distribution* of values as compared to what would be expected (e.g., salaries of different groups of employees compared to medians)
- *Order* or sequence of discrete events (e.g., are they really random?)

The number and variety of tests are considerable (see Siegel & Castellan, 1988, and Gibbons & Chakraborti, 1992, which have comprehensive coverage of tests), and Table 2 illustrates the roles of a variety in contrast with their parametric counterparts. If there is a need

to test whether a sample is typical of all samples that could be drawn from a population, the analogue to the *z*-test for ordinal data (ranks) is the KOLMOGOROV-SMIRNOV TEST, and for nominal data (frequencies for categories), it is the chi-square goodness-of-fit test. In both cases, population data are needed with which to compare the sample data. Similarly, there are analogues to the *t*-tests for two groups to test whether they belong to the same population. For ordinal data, the MANN-WHITNEY U-TEST will compare two unrelated groups, and for nominal data, the chi-square test for two groups will resolve the hypotheses. For related or matched groups, the corresponding test to the *t*-test for related groups is the Wilcoxon signed ranks test, and for nominal data, it is MCNEMAR'S CHI-SQUARE TEST (sometimes referred to as the McNemar change test). As can be seen from Table 3, there are similar analogues for three and more groups as well.

DESCRIPTORS OF CORRELATION AND ASSOCIATION

The results of surveys of single groups often endeavor to describe correlations (continuous and ordinal data) or associations (nominal) between pairs of variables. The statistic that is most appropriate depends on the nature of the data and, in this case, how each of the variables is measured, as shown in Table 3. If they are both the same (e.g., both are continuous), then it is a straightforward decision: the Pearson product moment correlation. But if they are not the same, then the choice is determined by the *lowest*. For example, if one variable was age (continuous) and the second was social class (ordinal), then the age data would have to be considered ordinal (ages changed to ranks) and the Spearman's rho would be most appropriate (see Table 3).

AN EXAMPLE OF A COMPARISON BETWEEN PARAMETRIC AND NONPARAMETRIC TESTS

Although the type of data is the ultimate determining factor in the choice between parametric and nonparametric tests, it is worth considering the types of research questions nonparametric tests can resolve and what the differences are between these and the ones answered by parametric tests. Sometimes, the questions may have to be changed because of the nature of the data that can be collected. To illustrate the issue, an example will be presented contrasting the use of parametric and nonparametric tests.

Two common tests will be used in an example to illustrate the basic differences: the t -test and chi-square or χ^2 . The latter is used primarily with nominal data, but is sometimes used with ordinal data where frequencies of ranked categories are used, due to its simplicity, although it may increase slightly the probability of making a Type II error.

Imagine the situation in which the research question asked whether there was a difference between two groups in the conservativeness of their political views. If the study were to use, say, an instrument on strength of conservative political views and generate a score for each person, then the means scores of the two groups could be compared to see if there were a difference using the t -test. On the other hand, if one were to record which party each group said they voted for in a recent election, assuming one party were identifiably conservative, then the frequency of voters for each

party for each group constitutes the data. To resolve the difference in voting patterns for the two groups, the χ^2 -test would be appropriate. But note that two different questions are being answered:

- Is there a difference in voting patterns between the two groups?
- Is there a difference in level of conservatism between the two groups?

Thus, one must be aware that the type of data collected may influence what question is being answered. In both cases, one might wish to infer whether there is a difference in conservatism between the two groups, but only the first really answers that question. We will come back to a numerical example of this question, but let us first look at some examples of nonparametric tests that are parallel to the z -test and t -test.

WHICH TO USE: T -TEST OR CHI-SQUARE?

This question was raised at the beginning of the entry. As was noted, the most obvious criteria for deciding the answer to this question are based upon the type of data collected to answer a research question. For example, if the question were

Is there a difference in how conservative patrons of two English pubs are?

then we would have to look at what data were used to answer the question. Simply put, if the data were scores from a random sample of patrons that would meet the criteria in the left column of Table 1, then the parametric t -test would be appropriate. But if the sample of patrons was classified into categories from more to less conservative, and frequencies belonging to each category tallied, then the data would require the nonparametric chi-square test.

To be more specific, let us map out two answers to the question by elaborating on two ways the variable could be operationally defined. We could define "how conservative" by designing a purpose-made test that would give a score relative to the level of conservatism and then administer it to patrons. Alternatively, we could take the view that action speaks louder than words and simply ask them which party they voted for (or would vote for) in an election, where the three main British political parties (at least in the past) have espoused political philosophies from conservative to socialist. Following the two lines of logic results in

Table 4 An Example of Two Ways to Ostensibly Answer the Same Question: “Is There a Difference in How Conservative Patrons of Two Pubs Are?”

<i>Parametric</i>	<i>Nonparametric</i>
Are patrons of one pub more conservative than another?	Do the patrons of two pubs vote along different party lines?
Questionnaire (20 questions, 5 point scale) as a measure of “conservatism.” The higher the score, the more conservative a person is. Individual score range: 20–100.	How many voted for (or would vote for) each of the political parties represented in local elections (frequencies for each party).
Compare mean scores for each pub with a <i>t</i> -test.	Compare voting frequencies for the two pubs using chi-square.

two different tests, as outlined in Table 4, and, logically following, the two statistical tests.

We can even illustrate the results with hypothetical data to show what could be the kind of outcomes in each case (we could use a more sensitive test for ordinal data, but the results would be much the same). This is shown in Table 5, with sample data for each of the two approaches. This raises a question about the measurements of the variable “conservatism”: Are they really measuring the same thing with two different instruments? Or, could there be other components or factors that might influence the outcomes of either instrument that may distort their responses? Who is asking the question? This is where the arguments begin.

The researcher using the questionnaire on conservatism could maintain that the instrument is

independent of *who* is running for office, the argument being that personality may encourage or discourage voters regardless of their political affiliation, and the choice of parties may not validly reflect how conservative a voter is. Although it might be argued that Conservatives are most conservative, Liberal Democrats are in between, and Labour voters are least conservative, for some specific issues and candidates, such a distinction is not valid and these are not strictly in order. Yet the researcher asking patrons to indicate which party they will vote for claims that this avoids abstractions and gets them to commit their beliefs to action, focusing on party and ideological commitment. What becomes apparent is that the tests are definitely appropriate for the data collected, but there is an issue as to which is most appropriate to answer the research question. In other words, the greatest potential for disagreement lies with the choice of OPERATIONAL DEFINITION (i.e., instrument) as a logical extension of the research question, and *not* with the choice of statistical test!

HISTORY

The brief history of nonparametric tests is highlighted by the developers and publication dates of selected tests (drawn from references in Howell, 2002; Kerlinger & Lee, 2000; Siegel & Castellan, 1988). This starts with Karl Pearson, who developed the chi-square test around 1900; the chi-square test is sometimes referred to as Pearson’s chi-square. Milton Friedman, later best known as an economist, devised his two-way analysis of variance for ranked data for related samples in 1937. A. Kolmogorov developed tests of ordinal data for goodness of fit for single samples and for comparing two independent samples about 1941,

Table 5 Data and Results for the Example of Two Ways to Ostensibly Answer the Same Question: “Is There a Difference in Political Persuasion for Patrons of Two Pubs?”

	<i>Parametric</i>		<i>Nonparametric</i>		
	<i>Green Toad</i>	<i>Red Herring</i>	<i>Green Toad</i>	<i>Red Herring</i>	
<i>Mean</i>	78	68	<i>Conservative</i>	35	20
<i>Standard Deviation</i>	22	24	<i>Liberal Democrats</i>	33	36
<i>Number</i>	87	76	<i>Labour</i>	19	20
			<i>Total</i>	87	76
$t = 2.76$ ($p < .05$)			$\chi^2 = 3.52$ (n.s.)		
$\therefore$ Reject H_0 : There probably is a difference in level of conservatism.			$\therefore$ Accept H_0 : There is probably no difference in voting patterns.		

and N. V. Smirnov published a table for the test in 1948. Wilcoxon (in 1945) and Mann and Whitney (in 1947) proposed other tests of two independent samples for ordinal data that are equivalent. W. H. Kruskal and W. A. Wallis proposed their one-way analysis of variance test for ordinal data in 1952. Quinn McNemar outlined his sign test for two related nominal groups in 1969.

—Thomas R. Black

REFERENCES

- Black, T. R. (1999). *Doing quantitative research in the social sciences: An integrated approach to research design, measurement, and statistics*. London: Sage.
- Bradley, J. V. (1985). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice Hall.
- Gibbons, J. D., & Chakraborti, S. (1992). *Nonparametric statistical inference* (3rd ed.). New York: Marcel Dekker.
- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.
- Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research*. Ft. Worth, TX: Harcourt.
- Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

NONPARTICIPANT OBSERVATION

Observation-based research that is conducted without subjects' awareness of being studied is often known as nonparticipant, unobtrusive, or nonreactive research. Two forms of nonparticipant observation, disguised observation and NATURAL EXPERIMENTS, in which permission is not sought, were once acceptable but are no longer ethically viable because they violate the principles of INFORMED CONSENT. Other styles of unobtrusive methodology, however, remain in the repertoire of sociologists, cultural anthropologists, and social psychologists.

Behavior trace studies, ARCHIVAL RESEARCH, and CONTENT ANALYSIS are three commonly used unobtrusive methods, although they do not consistently require the researcher to be involved in on-site observation of actual behavior. Researchers who use UNOBTRUSIVE METHODS with a strong component of on-site observation tend to rely on the systematic recording of data, often with the use of precoded checklists. Because they cannot verify their interpretations by asking informants questions about their behaviors, they must be especially careful to record their data in the most objective way possible. They must also be sure to

report their findings in aggregated ways so as not to compromise inadvertently the privacy of individuals whose identity might be revealed. When such observations are carried out by multiple researchers at multiple sites, care must be taken to ensure that data are recorded in a consistent fashion.

Nonparticipant observation is especially important when the subject is one about which participants might not be forthcoming when asked direct questions (e.g., covert behavior, such as the ways adolescents flirt with one another, or behavior in which the researcher cannot legally or ethically participate, such as drug use). Nonparticipant observation has been used in a variety of ETHNOGRAPHIC studies, for example, to characterize nonverbal communications (i.e., to understand how people use spatial arrangements or body gestures) or to describe the ways in which people make purchasing choices (but not to analyze the reasons behind their decisions). For example, Price (1989) studied the behaviors of customers in two urban pharmacies in Ecuador. She wanted to find out if any prescription drugs were being bought without prescriptions, and she was able to determine, after observing more than 600 sales transactions, that 51% of purchases were for prescription drugs for which the buyer had no prescriptions. She also learned that in such cases, pharmacy customers bought capsules in very small quantities (two or three pills at a time). In addition, Price unexpectedly learned that pharmacists (and even untrained drugstore clerks) were major purveyors of medical advice, especially for lower-class customers, who typically had limited access to physicians.

Richard Lee's (1993) widely cited study of the !Kung San people of southern Africa focused on food sources and the composition of diet. He set out to measure the distance people walked from camp to the gathering or hunting sites. He also calculated the weight of nuts collected or animals killed per hour of labor, the average caloric intake from both plant and animal resources, the number of different foods consumed, and the social relationships of those who shared foods of different types. On the basis of these observations, Lee was able to formulate more intelligent questions about food, diet, and social networks for a later phase of research than would have been the case had he entered the field armed only with his outsider's preconceptions about the people's behaviors.

Despite the reluctance of nonparticipant observers to interact directly with those being observed, it is sometimes necessary to ask questions to elicit contextual information; such background material can

be collected before the actual observation begins. By the same token, some observational researchers have found that it is sometimes necessary to ask questions designed to clarify the observed action; such questions can be asked after the observation has ended. But in nonparticipant observation, the main part of the research process should flow without the researcher's intervention.

—Michael V. Angrosino

REFERENCES

- Lee, R. B. (1993). *The Dobe !Kung*. New York: Holt, Rinehart and Winston.
- Price, L. J. (1989). In the shadow of biomedicine: Self-medication in two Ecuadorian pharmacies. *Social Science and Medicine*, 28, 905–915.
- Sechrest, L. (Ed.). (1979). *Unobtrusive measures today*. San Francisco: Jossey-Bass.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally.

NONPROBABILITY SAMPLING

A nonprobability sample is one not based on random selection methods.

The purpose of sampling is to obtain an accurate representation of the population from which the SAMPLE is drawn, so that accurate inferences about the population can be made on the basis of it. The samples most likely to be representative are large and randomly drawn (see RANDOM SAMPLING)—every member of the population has a known, nonzero probability of being part of the sample. (In *simple* random samples, every member would have an equal probability of being chosen, but there are also other variants.)

Often, random sampling is not possible or too expensive. It requires a complete list of the population or some other way of guaranteeing that every member is considered, and selected units cannot be replaced if they prove to be unavailable or refuse to take part. Under these circumstances, survey organizations often resort to the QUOTA SAMPLE, which is a *nonprobability* method. Here, the representation of the population is not left to chance but built into the sampling pattern. Knowing that gender is a variable of interest, for example, and that the relevant population is equally divided between men and women, the survey design will require interviewers to select equal numbers of men

and women—they will be given a quota of men and women to find. Properly conducted, this kind of sample is necessarily representative of the population in terms of the quota variables; this is built into the design.

It may not be representative in other terms, however, because the interviewers are left free to exercise choice about how they fill the quota. In a survey that British Open University students carried out as part of their coursework, for example, every student was responsible for collecting four cases according to a quota system that took account of gender, age, and social class. (This yielded a sample of about 2,000 people nationally.) When we came to analyze the data, we found that this sample appeared representative of most of the known facts about the population. However, despite the national tendency for middle-class people of both genders to vote Conservative, in our sample the middle-class women were more likely to vote for the Labour Party. The explanation was almost certainly that when the students were looking for middle-class women, they tended to go to their children's schools and approach teachers—and research had shown that middle-class people in the public sector of the economy were, in fact, more likely than others to vote for the Labour Party.

Quota samples at least represent the population in certain respects, but other forms of nonprobability sample are less likely to do so (see CONVENIENCE SAMPLE). Much research is conducted on what may best be described as a *haphazard* sample—stopping people “at random” in the street, for example. Other studies are carried out on “samples of opportunity”—education research carried out on the class you happen to be teaching, for example. Other studies use *volunteer* samples—they advertise for participants. In all of these cases, it is very likely indeed that the representation of the population will be imperfect. We shall fail to contact the kind of people who would not be on the street at that time. Our class may be untypical of the school, which in turn may not be representative of the total population of the town (which in turn may not be typical of the country as a whole). Volunteer samples cannot reach those who do not volunteer, and volunteers are most likely to include people with strong views on the advertised subject and less likely to include those who are indifferent to it.

Information based on nonprobability samples is therefore to be treated with caution. Alternatively, the researcher needs to think carefully about what population *is* represented. A survey on loneliness with a volunteer sample drawn by advertising in

coffee bars, for example, will not represent true social isolates—who do not go out to coffee bars—and it will not represent people too shy to respond to the advertisement. Within those limitations, however, it may provide useful information.

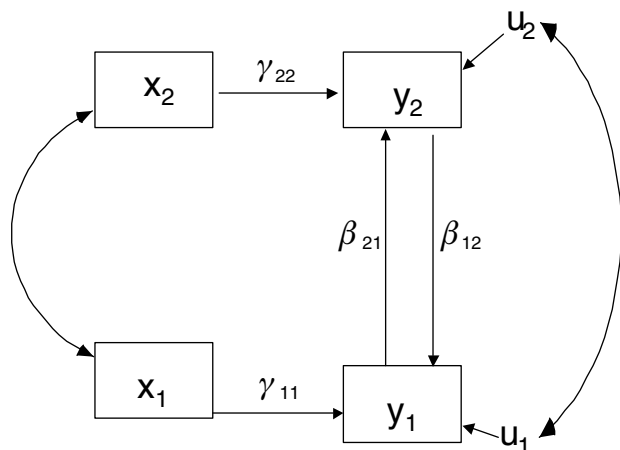
One point of view is that the *achieved* sample is nonrandom in *all* surveys, however rigorous the design. The sampling pattern may have been random, but the people whom we fail to reach or who refuse to participate are very unlikely to be a random subset of the planned sample; they will differ from those who do respond in systematic ways.

—Roger Sapsford

NONRECURSIVE

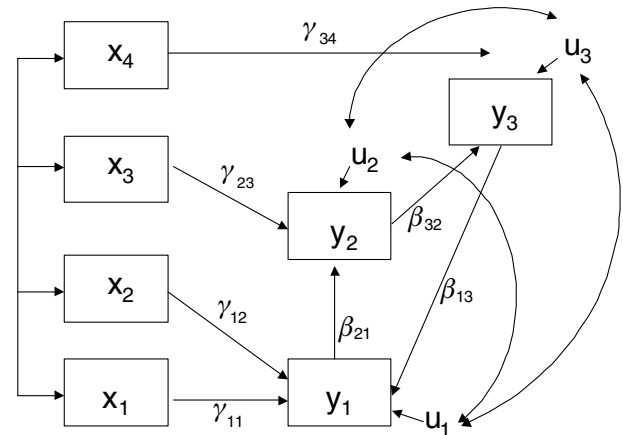
Many social scientific theories specify MODELS of more than one equation. A nonrecursive system of equations is a type of SIMULTANEOUS EQUATION system where the equations are linked through RECIPROCAL RELATIONSHIPS between dependent variables, or correlations between errors in the equations. In nonrecursive models, either (a) there is reciprocal causation or feedback loops, and/or (b) the errors in the equations are correlated. The presence of either reciprocal causation or correlated errors (or both) indicates that a model is nonrecursive. In contrast, if there is neither reciprocal causation nor correlated errors, a simultaneous equation system is called RECURSIVE. SEEMINGLY UNRELATED REGRESSION models are a special case of nonrecursive models where the equations are related only through correlations between the errors.

A simple nonrecursive model with reciprocal causation can be represented by the following PATH DIAGRAM:



Here, y_1 affects y_2 while y_2 in turn affects y_1 . Therefore, a change in y_1 feeds back upon itself. Notice also that the errors in the equations are correlated.

A nonrecursive model with a feedback loop could appear as follows:



Here, y_1 affects y_2 , y_2 affects y_3 , and y_3 returns to affect y_1 . Again, a change in y_1 will feed back upon itself.

Consider the two-equation nonrecursive system of equations corresponding to the first diagram:

$$y_1 = \beta_{12}y_2 + \gamma_{11}x_1 + u_1$$

$$y_2 = \beta_{21}y_1 + \gamma_{22}x_2 + u_2$$

where $\text{Cov}(u_1, u_2) \neq 0$.

In the system of equations, ENDOGENOUS VARIABLES are denoted with y s, while EXOGENOUS VARIABLES are denoted with x s. β s indicate the effect of an endogenous variable on another endogenous variable, whereas γ s indicate the effect of an exogenous variable on an endogenous variable. The u s indicate errors in the equations. In order for the system to be in equilibrium, $|\beta_{12}\beta_{21}| < 1$.

Nonrecursive systems of equations must avoid the IDENTIFICATION PROBLEM. That is, whether the parameters of the model are estimable must be established. Many methods to identify nonrecursive models are possible, including the order condition and the rank condition (see Bollen, 1989, pp. 98–103). Rigdon (1995) provides a straightforward way to identify many types of nonrecursive models.

Estimation of nonrecursive systems of equations requires that estimation take account of the information from other equations in the system. Therefore, if ORDINARY LEAST SQUARES (OLS) regression is applied to a nonrecursive simultaneous equation system, the estimates are likely to be biased and inconsistent. Instead,

nonrecursive models can be estimated by two-stage least squares, three-stage least squares, and MAXIMUM LIKELIHOOD ESTIMATION. It is possible to view all of these estimators as INSTRUMENTAL VARIABLE estimators.

—Pamela Paxton

REFERENCES

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Greene, W. H. (1997). *Econometric analysis* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Rigdon, E. E. (1995). A necessary and sufficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research*, 30, 359–383.
- Heise, D. R. (1975). *Causal analysis*. New York: Wiley.

NONRESPONSE

There are two broad categories of nonresponse in survey research: unit nonresponse, when a subject does not complete the survey instrument, and item nonresponse, when a respondent fails to answer one or more of the questions posed.

UNIT NONRESPONSE

Generally speaking, response rates have been declining in surveys of households and establishments over the past few decades. This nonresponse contributes to increased costs of data collection and may introduce nonresponse BIAS.

Groves and Couper (1998) have created a typology of exogenous and endogenous factors influencing survey cooperation. Exogenous factors, which are largely out of the control of the survey researcher, include the social environment (e.g., level of urbanicity in the area to be surveyed or the amount of advance publicity about the survey) and sample members' social psychological attributes and levels of knowledge about the topic of interest. Endogenous factors include the data collection protocols (mode of administration, incentives, length of the instrument, etc.) and, in interviewer-administered surveys, interviewer selection and training.

It can be challenging to enlist cooperation in surveys of establishments. In many cases, one individual does not have all the information necessary to respond. In

telephone studies, it can be difficult for interviewers to get past gatekeepers to speak directly with respondents. Some companies have "blanket" policies that prohibit survey participation by their employees. Even in establishments without such a policy, the person receiving the invitation to participate may have to get authorization to participate from someone else in the organization.

There are research design features that have demonstrated positive effects on response rates. The numbers and types of contacts with sample members are important factors in enlisting cooperation. Mail surveys often include postcard reminders, replacement mailings of questionnaire packets, and telephone reminder calls to nonresponders. In TELEPHONE SURVEYS, the timing of calls is important to increasing the odds of finding the potential participant at home. COMPUTER-ASSISTED DATA COLLECTION systems can incorporate rules to make sure that calls are attempted during daytime and evening hours, on weekdays and weekends, and that calls are placed in different weeks. Other design features that have been shown to influence response include the length of the field period (generally the longer, the better); the survey sponsor; advance letters that provide information about the study and let participants know that an interviewer will be calling; and monetary and nonmonetary incentives. Prepaid cash incentives have proven to be the most productive type of incentive. Refusal conversions and increasing the numbers of call attempts also tend to increase response rates.

Information on nonrespondents is often limited. In nonresponse studies, a random sample of nonrespondents is selected, and intensive recruitment efforts are undertaken (e.g., offering or increasing incentives, offering different modes of administration, and/or increasing the numbers of contacts). Usually, these surveys include a few key questions from the survey and demographic questions. The data collected can be used to characterize nonrespondents and make postsurvey adjustments to the data.

When respondents are systematically different from nonrespondents in ways that are relevant to the research questions being addressed by the survey, NONRESPONSE BIAS occurs. Nonresponse is widely acknowledged as a major source of ERROR in survey research. Raising response rates is desirable in itself because it increases the credibility of data. However, the most important result is that improving response rates also tends to reduce nonresponse bias. The effects of additional efforts to reduce nonresponse bias depend on the

size of the difference between respondents and nonrespondents. The substantive significance of non-response reduction will inevitably vary by topic, population, and methods of data collection.

ITEM NONRESPONSE

Respondents can fail to follow skip patterns accurately, intentionally omit an answer to a question they are uncomfortable answering, or stop answering survey items before the questionnaire is completed. This item nonresponse is more of an issue when surveys are self-administered than when an interviewer is present. Interviewers develop rapport with respondents and help them negotiate the instrument, and they can repeat questions or provide standardized definitions for unfamiliar terms when necessary.

Missing answers are more likely in questions about sensitive issues than in less threatening items. Questions concerning income levels and socially undesirable behaviors, such as recreational drug use or criminal behavior, often have MISSING DATA. Items that ask for the respondent's income often have the highest level of item nonresponse in a given survey. Asking about income in a series of nested questions and/or providing income ranges as response options can help to improve the quality of these data.

Poorly designed questions (e.g., ambiguous or double-barreled items) often yield missing data, either unit or item. Items that increase respondent burden by requiring extra effort on the part of the respondent also increase the likelihood of nonresponse. Examples are questions that ask respondents to answer in percentages that sum to 100%, items with long recall periods, and open-ended items in self-administered questionnaires. During instrument development, the testing of candidate items in cognitive interviews with people who mirror key characteristics of the population to be surveyed and in field PRETESTS will help identify problematic questions. These items can then be reworked before the survey is fielded.

Skip instructions embedded in an instrument indicate what question the respondent should answer next based on the answer to the trigger item. Faulty skip patterns can occur when respondents fail to answer a question they should, and when they answer items that should have been skipped. Well-designed instruments will minimize skip errors. Survey researchers need to carefully consider the content of both visual and

textual information about skip instructions presented to respondents.

The issue of the differences between “purely” missing answers—items about which respondents have not formulated an opinion and items that are not applicable to a respondent—also should be considered when thinking about item nonresponse. The use of screening questions is recommended. These “trigger” questions are used to identify the appropriate denominator for substantive questions by allowing respondents who do not have relevant experience to inform their answers to factual items or those who do not have a position on opinion items to skip the “target” item. There is, however, concern that offering a no-opinion option may give respondents an easy way out of responding to questions where it is reasonably certain they have an opinion.

Item nonresponse in establishment surveys can result when respondents are asked to provide information the company either does not collect or collects in units that do not fit the response choices offered. Firms also consider some information proprietary—employees may be prohibited from releasing certain requested information.

Research design that maximizes response rates and instrument development that optimizes question design are pivotal to reducing nonresponse in surveys.

—Patricia M. Gallagher

REFERENCES

- Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: Sage.
- Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.
- Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.). (2002). *Survey nonresponse*. New York: Wiley.
- Schwarz, N., & Sudman, S. (Eds.). (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. San Francisco: Jossey-Bass.

NONRESPONSE BIAS

NONRESPONSE occurs when a study fails to gather data from all the elements that have been sampled from the frame. This can occur in any form of research, not just in sample surveys. Measurement of the size of the nonresponse is the response rate, and it can be

computed in several informative ways (cf. AAPOR, 2000). Although response rates provide useful information about the study's efficiency in measuring all sampled elements, they do not directly provide any information about whether the data have any nonresponse bias. That is, "low" response rates do not automatically mean that nonresponse bias is present (cf. Curtin, Presser, & Singer, 2000; Merkle & Edelman, 2002). However, whenever nonresponding elements differ from the responding elements on the measures of interest, then nonresponse error occurs, most typically in the form of bias. Nonresponse bias is a function of *both* the size of the nonresponse and the size of the differences between responders and nonresponders (e.g., Groves, 1989).

If responders and nonresponders are not different on measures of interest, then there will be no nonresponse bias, regardless of the study's response rate. Nonresponse bias can occur at the unit level (e.g., person) or at the item level (e.g., a survey question). Many methodological techniques have been devised to try to reduce the likelihood of nonresponse, which in surveys occurs most often because of noncontacts, refusals, and language barriers. Many statistical WEIGHTING and IMPUTATION techniques have been devised to adjust for possible nonresponse bias at both the unit and item levels. However, typically no information is available from the nonresponding elements; thus, there is no guarantee that these statistical adjustments will, in fact, reduce any nonresponse bias that may be present. Special follow-up studies of nonresponders can gather information that can be used to determine how the nonresponders in the original study differed from responders.

—Paul J. Lavrakas

REFERENCES

- American Association for Public Opinion Research (AAPOR). (2000). *Standard definitions: Final dispositions for case codes and outcome rates for surveys*. Ann Arbor, MI: Author.
- Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64, 413–428.
- Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
- Merkle, D. M., & Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 243–258). New York: Wiley.

NONSAMPLING ERROR

There are many forms of potential error in social research that a prudent researcher will (a) try to minimize and/or (b) measure (cf. Groves, 1989). Error is due to the various causes of BIAS (i.e., constant error) and VARIANCE (i.e., variable error) that can occur whenever data are gathered, regardless of whether the research is QUANTITATIVE RESEARCH or QUALITATIVE RESEARCH. SAMPLING ERROR is the variance associated with any research sample that is not a census of the population of interest. *Nonsampling error* is the term that traditionally has been used to refer to all other sources of bias and variance, and it includes coverage error, nonresponse error, and measurement error.

Coverage error refers to the bias that may result when the frame (the list) used to represent the population of interest fails to adequately cover (include) the entire population. *NONRESPONSE error* refers to the bias and variance that may result when data are not gathered from all of the elements sampled from the frame. *Measurement error* refers to potential bias and variance that can be caused by any of the following: (a) the person who gathers the data (e.g., an interviewer, observer, or coder); (b) the instrument that is used to gather the data (e.g., a questionnaire, observational or coding form); (c) the subject/respondent being measured; (d) the mode of data collection (e.g., in-person vs. self-administered). The literature abounds with methodological and statistical studies that address various techniques that can be used to try to reduce the possibility that nonsampling errors will occur and/or ways to measure and adjust for such errors when they do result. Of note, many think that these forms of errors apply only to sample surveys, when, in fact, each type of error has its counterpart in any form of social research, including OBSERVATIONAL RESEARCH, CONTENT ANALYSIS, and all forms of qualitative research.

—Paul J. Lavrakas

REFERENCE

- Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

NORMAL DISTRIBUTION

The normal distribution (sometimes referred to as the Gaussian distribution) has been applied to problems in the social sciences many times. Many continuous random variables can be modeled as a normal PROBABILITY distribution whose density function produces a symmetric, bell-shaped curve. The normal density function for a continuous random variable Y is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}},$$

where $-\infty < y < \infty$, $-\infty < \mu < \infty$, and $\sigma > 0$, and μ and σ are the POPULATION mean and STANDARD DEVIATION, respectively, such that $E(Y) = \mu$, and $V(Y) = \sigma^2$. Notice that these are the only two PARAMETERS that define the density of the normal distribution.

The normal distribution has a special property that holds across parameter values: 68% of the distribution lies within 1 standard deviation above and below the mean, 95% of the distribution lies within 2 standard deviations above and below the mean, and 99.7% of the distribution lies within 3 standard deviations of the mean. However, because the distribution lies between $-\infty$ and ∞ , a normally distributed random variable

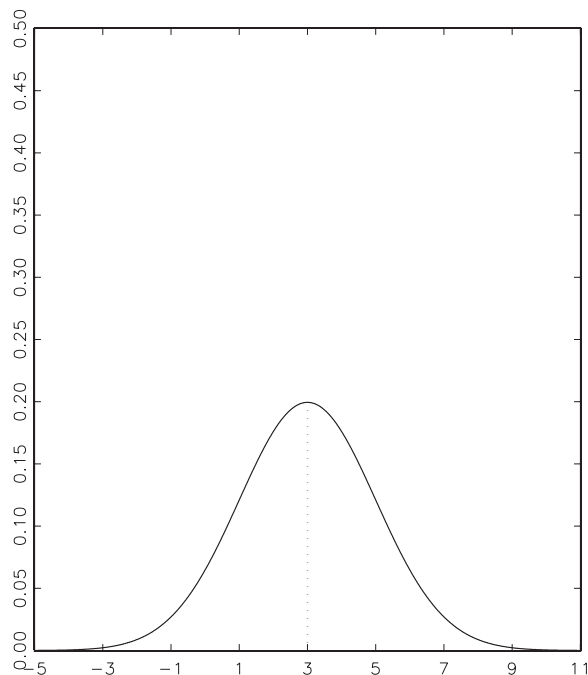


Figure 1 $Y_i \sim \text{Normal}, \mu = 3, \sigma = 2$

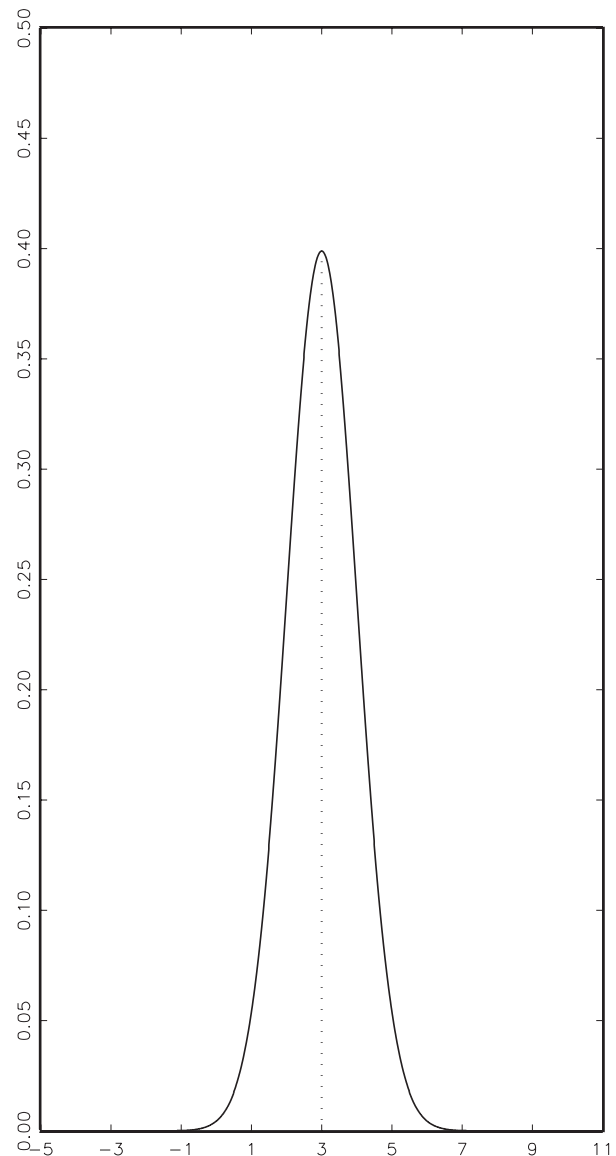


Figure 2 $Y_i \sim \text{Normal}, \mu = 3, \sigma = 1$

has events with nonzero probability occurring anywhere on the real-number line. In addition, because the distribution is symmetric and has a unique maximum, its MEAN, MEDIAN, and MODE are the same. Furthermore, the two inflection points are located 1 standard deviation away from the mean. Figures 1 and 2 demonstrate these points. Note that the two distributions are both centered around 3, but the spread of the first is greater than that of the second. However, a very high or very low number is a possible event (even with very low probability) in both distributions.

A major reason for the centrality of the normal distribution in the social sciences is the CENTRAL LIMIT THEOREM. The theorem links any probability function (as long as its μ and σ are finite) to the normal distribution by proving that regardless of the distribution of the original variable, the mean and the sum of the variable are normally distributed for large enough SAMPLES (or, more strictly, as n approaches infinity). A particular application of the theorem is the normal approximation to the BINOMIAL distribution. The application establishes that a fraction of successes in a series of n BERNOULLI trials with a probability of success in each one, p , is approximately normally distributed with a mean p and variance $\frac{p(1-p)}{n}$.

In addition, following the ASSUMPTION that the ERROR terms in the frequently used LEAST SQUARES REGRESSION are distributed normally, the least squares estimators are distributed normally, a property that helps with many questions of STATISTICAL INFERENCE.

THE STANDARD NORMAL DISTRIBUTION

To compare values of two normally distributed variables, each drawn from a different normal distribution, a standardization is required. The standardization shifts the means of the distributions and adjusts their variances to comparable grounds, such that a comparison of, say, a grade in Advanced Akkadian and a grade in Introduction to Astrophysics is made possible.

The standardization procedure is $Z_i = \frac{Y_i - \mu}{\sigma}$, such that each value Z_i represents the difference from Y_i to its mean in units of its standard deviation (σ) of the original variable Y_i . Therefore, the mean value of Z_i must be zero, and the standard deviation is equal to 1.

Thus, the density function for the standard normal distribution is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Evaluating the probability of a normal variable being between points a and b [$P(a \leq Y_i \leq b)$] requires evaluation of the area under the normal density function between points a and b , or, in other words, evaluation of the integral

$$\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy.$$

Unfortunately, a closed-form expression for the integral does not exist, hence, numerical techniques are used, and in practice, statistical tables that summarize the density are often put to use. Because the distribution

is symmetric around μ , symmetric areas need to be calculated only on one side of the mean. Otherwise, for convenience, areas to the left of the mean are calculated as if they were to the right of the mean. Figure 3 demonstrates this point with respect to the standard normal distribution.

$$\begin{aligned} P(a < Y_i < b) &= P(Y_i < b) - P(Y_i < a) \\ &= P(Y_i < b) - P(Y_i > -a) \\ &= P(Y_i < b) - [1 - P(Y_i < -a)] \\ &= P(Y_i < b) + P(Y_i < -a) - 1. \end{aligned}$$

Suppose $a = -1.5$ and $b = 0.8$. Then

$$\begin{aligned} P(-1.5 < Y_i < 0.8) &= P(Y_i < 0.8) \\ &\quad + P(Y_i < 1.5) - 1 = 0.7881 \\ &\quad + 0.9332 - 1 = 0.7213. \end{aligned}$$

LINEAR TRANSFORMATION

Often, social scientists would work with a variable that is a LINEAR TRANSFORMATION of an original variable of interest. For example, a researcher might measure income in dollars or thousands of dollars, or distance in kilometers or miles. When the original variable is distributed normally, the linear transformation of it is also distributed normally. Furthermore, if we know the mean and standard deviation of the original variable, it is easy to extract the parameter values of the transformed variable when they are linearly related. If $Y_i \sim N$ such that $E(Y) = \mu$ and $V(Y) = \sigma^2$, and X is a linear transformation of Y such that $X_i = aY_i + b$, then $E(X_i) = E(aY_i + b) = aE(Y_i) + b = a\mu + b$, and $V(X_i) = V(aY_i + b) = a^2V(Y_i) = a^2\sigma^2$.

MULTIVARIATE NORMAL DISTRIBUTION

If, rather than a single variable, Y_i is a vector of N random variables normally distributed, we need a multivariate distribution to model this vector. The multivariate normal distribution with N random variables can be written as a function of an $N \times 1$ vector μ and an $N \times N$ variance-covariance matrix Σ , such that

$$f(y) = \frac{1}{\sqrt{|\Sigma|}(2\pi)^{-\frac{N}{2}}} e^{-\frac{1}{2}[(y_i - \mu)\Sigma^{-1}(y_i - \mu)]}.$$

When Y_i includes only one random variable, this expression reduces to the univariate normal density, and its variance-covariance matrix is simply σ^2 .

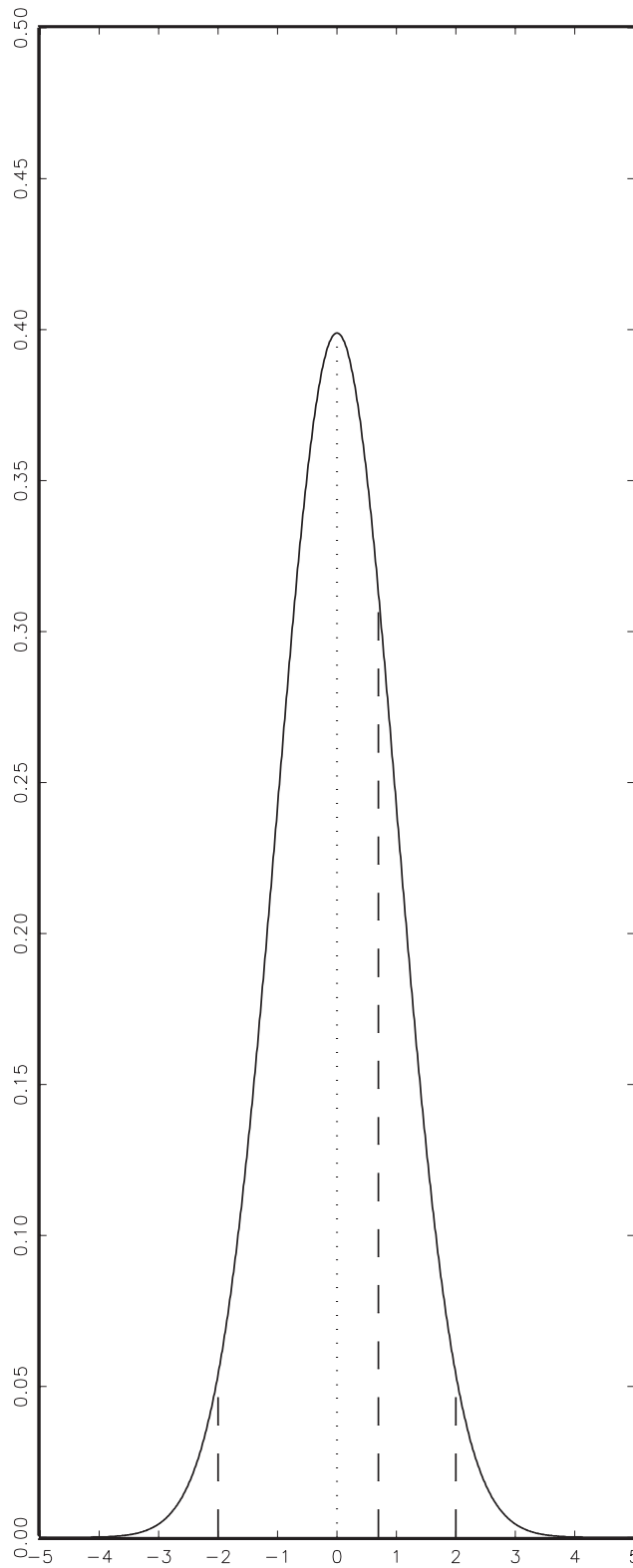


Figure 3 $Y_i \sim \text{Standard Normal } (\mu = 0, \sigma = 1)$

When the variables are independent, Σ is diagonal (i.e., the off-diagonal entries are equal to zero). In this special case, the product of the univariate normal distributions will equal the multivariate normal distribution.

LOG-NORMAL DISTRIBUTION

A distribution that might be helpful to describe variables that can take only a non-negative value (such as income or weight) is the log-normal distribution. If Z_i is distributed standard normal such that $Z_i = \frac{\ln Y_i - \mu}{\sigma}$, then Y_i is distributed log-normal with mean μ and variance σ^2 , such that $Y_i = e^{\sigma Z_i + \mu}$.

HISTORICAL DEVELOPMENT

The normal distribution is often associated with Carl Friedrich Gauss (1777–1855). In a book published in 1809, Gauss associated the normal density with the method of least squares. Yet Gauss cites work by Laplace from 1774, where he touches on the distribution, and indeed, some works refer to the distribution as Gauss-Laplace. However, historical scholarship traces the distribution's origin back even further to a publication by De Moivre from 1733 (Stigler, 1999, pp. 284–285).

Indeed, the idea that a normal distribution can describe statistical processes was developed in the 18th century. For much of the 19th century, they were called “error curves” because they were first used to describe the distribution of measurement errors. It was not until the 1870s that the phrase “normal distribution” was introduced.

—Orit Kedar

REFERENCES

- Greene, W. H. (1993). *Econometric analysis*. Englewood Cliffs, NJ: Prentice Hall.
- King, G. (1989). *Unifying political methodology*. New York: Cambridge University Press.
- Stigler, S. M. (1986). *The history of statistics*. Cambridge, MA: Harvard University Press.
- Stigler, S. M. (1999). *Statistics on the table*. Cambridge, MA: Harvard University Press.
- Wackerly, D. D., Mendelhall, W., III, & Scheaffer, R. L. (1996). *Mathematical statistics with applications* (5th ed.). Belmont, CA: Duxbury.

NORMALIZATION

In many popular statistical models, we assume that some component of a variable Y has a NORMAL DISTRIBUTION. For example, in the LINEAR REGRESSION model $Y = \alpha + \beta X + \varepsilon$, we typically assume that the error term ε is normal. Although minor departures from normality may be acceptable, distributions with heavier-than-normal tails can compromise statistical estimates. In such cases, it may be preferable to transform Y so that the pertinent component is closer to normality. Transforming a variable in this way is called *normalization*.

If the pertinent component of Y has one heavy tail (SKEWED), then we often apply a *power transformation*. True to their name, power transformations raise Y to some power p (i.e., they transform Y into Y^p). Powers greater than 1 reduce *negative* skew: An example is the quadratic transformation Y^2 ($p = 2$). Powers between 0 and 1 reduce *positive* skew: An example is the square-root transformation or $\sqrt{Y}$ or $\sqrt{Y + 1/2}$ ($p = 0.5$), which is common when Y represents counts or frequencies. For a power of 0, the power transformation is defined to be $\log(Y)$, which reduces positive skew in much the same way as a very small power. Negative powers have the same effect as positive powers applied to the reciprocal $\frac{1}{Y}$ and are used when the reciprocal has a natural interpretation—as when Y is a rate (events per unit time), so that $\frac{1}{Y}$ is the time between events.

In sum, the family of power transformations can be written as follows:

$$t(Y; p) = \begin{cases} Y^p & \text{if } p \neq 0, \\ \ln(Y) & \text{if } p = 0. \end{cases}$$

Power transformations assume that Y is positive; if Y can be zero or negative, we commonly make Y positive by adding a constant. There are formal procedures for estimating the best constant to add, as well as the power p that yields the best approximation to normality (Box & Cox, 1964). However, the optimal power and additive constant are usually treated only as rough guidelines.

If the pertinent component of Y has *two* heavy tails (excess KURTOSIS), we may use a *modulus transformation* (John & Draper, 1980),

$$t(Y; p) = \begin{cases} \text{sign}(Y)[\frac{(|Y|+1)^p - 1}{p}] & \text{if } p \neq 0, \\ \text{sign}(Y)[\ln(Y) + 1] & \text{if } p = 0, \end{cases}$$

which is a modified power transformation applied to each tail separately. Non-negative powers p less than 1 reduce kurtosis, while powers greater than 1 increase kurtosis. Again, there are formal procedures for estimating the optimal power p (John & Draper, 1980). If Y is symmetric around 0, then the modulus transformation will change the kurtosis without introducing skew. If Y is not centered at 0, it may be advisable to add a constant before applying the modulus transformation.

Other normalizations are typically used if Y represents proportions between 0 and 1: the arcsine or angular transformation $\sin^{-1}(\sqrt{Y})$, the logit or logistic transformation $\ln(\frac{Y}{1-Y})$, and the probit transformation $\Phi^{-1}(Y)$, where Φ^{-1} is the inverse of the cumulative standard normal density. The logit and probit are better normalizations than the arcsine, which is gradually disappearing from practice.

Even the best transformation may not provide an adequate approximation to normality. Moreover, a transformed variable may be hard to interpret, and conclusions drawn from it may not apply to the original, untransformed variable (Levin, Liukkonen, & Levine, 1996). Fortunately, modern researchers often have good alternatives to normalization. When working with non-normal data, we can use a GENERALIZED LINEAR MODEL that assumes a different type of distribution. Or we can make weaker assumptions by using statistics that are “distribution-free” or nonparametric.

In addition to the definition given here, the word *normalization* is sometimes used when a variable is STANDARDIZED. It is also used when constraints are imposed to ensure that a system of SIMULTANEOUS EQUATIONS is identified (e.g., Greene, 1997).

—Paul T. von Hippel

REFERENCES

- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society (B)*, 26(2), 211–252.
- Cook, R. D., & Weisberg, S. (1996). *Applied regression including computing and graphics*. New York: Wiley.
- Greene, W. H. (1997). *Econometric analysis* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- John, N. R., & Draper, J. A. (1980). An alternative family of transformations. *Applied Statistics*, 29(2), 190–197.
- Levin, A., Liukkonen, J., & Levine, D. W. (1996). Equivalent inference using transformations. *Communications in Statistics, Theory and Methods*, 25(5), 1059–1072.

Yeo, I.-K., & Johnson, R. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87, 954–959.

NUD*IST

A software program designed for computer-assisted qualitative data analysis. As such, it allows qualitative data to be coded and retrieved. It also supports GROUNDED THEORY practices, such as the generation of MEMOS. The name is an abbreviation of Non-numerical Unstructured Data Indexing, Searching, and Theorizing.

—Alan Bryman

NUISANCE PARAMETERS

A nuisance parameter is a type of parameter that is of no or only secondary interest in parametric ESTIMATION. For example, in a certain statistical model that involves a normally distributed random variable, the researcher's interest often focuses on the estimation of the expectation of the variable μ , whereas the variance of the random variable σ^2 is of no interest and is then considered a nuisance parameter.

—Tim Futing Liao

NULL HYPOTHESIS

The null hypothesis is an essential component of HYPOTHESIS TESTING in social research. It is relevant when quantitative measures of social activities have been made and when hypotheses derived from theories are to be tested. The theoretical (or research) hypothesis defines a certain pattern to be found in the data, and the statistical analysis is designed to evaluate the extent to which the evidence from the collected measurements supports the existence of that pattern. The null hypothesis is the hypothesis that the pattern of data found has occurred simply by chance.

Take an example of a RANDOM SAMPLING of participants drawn from the human POPULATION, which is randomly divided into two groups. One group may receive a treatment, such as a small dose of alcohol,

whereas the other receives a placebo (to conceal from the participants which group they are in). Their performance on a driving task (or other related skill) may then be measured. The research hypothesis will be that the two groups will have different levels of performance. The null hypothesis is that the alcohol has no effect and any resulting difference between the two groups is due to random chance. If the null hypothesis is true, we expect the mean performance of the two groups to be the same. However, any RANDOM ASSIGNMENT of participants into two groups almost always will produce two groups whose measured performance differs by some amount, even when both groups are treated exactly the same. However, the larger the difference in the means, the less probable is such an outcome. Hypothesis testing relies on estimating the probability (or chance) of obtaining the results we measure.

If it is assumed that the null hypothesis is true (in this example, that the alcohol has no effect), it is possible to calculate the chance of getting any pattern of data that may result from the measurement of two randomly allocated groups of participants. Usually, this is done by using this assumption (that the null hypothesis is true) to calculate (or otherwise estimate) the SAMPLING DISTRIBUTION of a statistic (such as *T* RATIO or CHI-SQUARE). This distribution is then used to compare with the sample statistic calculated from the measurements taken.

In principle, even if there is no effect of the treatment (so the null hypothesis is true), any mathematically possible pattern in the data may be found by chance alone, however unlikely. The smaller the probability of finding a given pattern of results by chance, the more we consider that we have evidence that supports the research hypothesis. Some statistical methods set a conventional probability level, below which the procedure allows the decision to be made that the null hypothesis may be rejected and the research hypothesis accepted. Other methods consider that there is no hard-and-fast decision rule that can be universally applied, and that the evidence obtained with one set of measurements can be fully evaluated only in the context of other research that has investigated similar problems in similar ways (see, e.g., MacDonald, 2002).

—Martin Le Voi

See also SIGNIFICANCE TESTING, *P* VALUES, HYPOTHESIS.

REFERENCE

MacDonald, R. R. (2002). The incompleteness of probability models and the resultant implications for theories of statistical inference. *Understanding Statistics*, 1(3), 167–189.

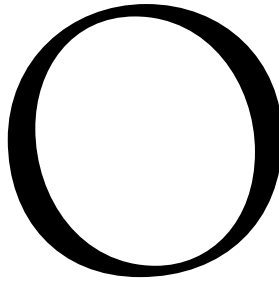
NUMERIC VARIABLE. *See*
METRIC VARIABLE

NVIVO

NVivo or NUD*IST Vivo is a software program designed for computer-assisted qualitative data

analysis. As such, it allows qualitative data to be coded and retrieved. It also supports GROUNDED THEORY practices, such as the generation of MEMOS. It is a variant (but not a new release) of NUD*IST but is often regarded as more suited to the fine-grained analysis that is particularly popular among practitioners of QUALITATIVE RESEARCH.

—Alan Bryman



OBJECTIVISM

Objectivism is the view that there can be a “permanent, ahistorical matrix or framework to which we can ultimately appeal in determining the nature of rationality, knowledge, truth, reality, goodness or rightness” (Bernstein, 1983, p. 8). Although long challenged by the phenomenological and hermeneutic traditions in social science, it was nevertheless the default position in both the natural and social sciences until the 1960s. Thomas Kuhn’s *The Structure of Scientific Revolutions*, first published in 1962 (Kuhn, 1970), challenged this orthodoxy in natural science (and by extension in social science) by claiming that there is no linear progress in science; rather, there are “paradigms” of thought and practice that are taken to represent scientific truth at particular times. These are subject to occasional and dramatic revolution that can be accounted for more readily by social and psychological factors than by objective facts about the world. Indeed, so opposed are competing paradigms that they are epistemologically incommensurate.

Kuhn’s work influenced a number of antiobjectivist positions in social science (or perhaps more correctly social studies, for in these positions there was often a denial the social world could be studied scientifically): RELATIVISM, social CONSTRUCTIONISM, and subjectivism. Though sometimes an antiobjectivist view will incorporate all these oppositions, they are not necessarily synonymous. Relativists will deny that knowledge and/or morality can be judged from a perspective that is not embedded in subjective epistemological or moral positions, and thus they claim

there can be no privileging of any one truth or morality. Subjectivists will deny the existence of an objective reality existing outside of agents’ perceptions or constructions of it. Social constructionism, though similar, places stress on the social construction of what we call “reality.”

Objectivism, like so many other “isms,” is often something of a straw person, and there are few now who would wholeheartedly embrace either it or its complete denial. There are, however, more moderate successor positions in respect of objectivity as an achievable goal or REALISM as an ontological stance in social theory and methodology. In the first case, the question addressed is to what extent can objectivity, as a pursuit of truth in the social world, be retained when the social world is actively constructed by its participants? Realists have attempted to find a middle ground by accepting that the social world is socially constructed, but once it comes into being it exists as an objective reality independent of any individual conception of it.

One influential attempt to transcend objectivism and its oppositions came from Pierre Bourdieu (1977), who associated it with the STRUCTURALISM of Durkheim, Althusser, and Lévi-Strauss and the consequent “disappearance” of the agent from social theory. Conversely, he saw the ETHNOMETHODOLOGISTS and SYMBOLIC INTERACTIONISTS as overly concentrating on agents’ subjectivity and ignoring the objective structures acting upon them. Instead, he proposed that there existed a dialectic relationship between subjective and objective factors in the practice of the construction of social reality. Objective structures bear upon day-to-day interactions, and although agents will reproduce and transform them, the ways in which they are presented

will make a difference in how they will be reproduced and transformed.

—Malcolm Williams

REFERENCES

- Bernstein, R. (1983). *Beyond objectivism and relativism: Science, hermeneutics and praxis*. Oxford, UK: Basil Blackwell.
- Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge, UK: Cambridge University Press.
- Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press. (Original work published 1962.)

OBJECTIVITY

The central meaning of the term “objectivity” concerns whether inquiry is pursued in a way that maximizes the chances that the conclusions reached will be true. As will become clear, however, the word has a number of other related meanings.

The fortunes of this concept underwent a dramatic change in the second half of the 20th century. It had been treated by most researchers as identifying an essential requirement of research, but it came to be regarded by some as little more than an ideological disguise for the interests of scientists, especially those who are male, white, and from Western countries.

The various meanings of “objectivity” can be clarified by distinguishing its adjectival and adverbial forms. Confusion often arises from conflating these. First, “objective” can be applied as an adjective to phenomena in the world. In this sense, it is sometimes taken to imply that a thing exists, rather than being a mere appearance or figment of imagination. Alternatively, “objectivity” can mean that a thing belongs to the “external” world, not to the “internal,” psychological world of a subject. In this sense, whereas tables and chairs are objective, thoughts and feelings are not, even though they both may be real rather than mere appearances.

“Objective” also can be applied as an adjective to research findings, and here it is equivalent in meaning to “true.” This is what is sometimes referred to as “ontological objectivity” (Newell, 1986). However, given the equivalence, this usage seems superfluous and in most cases is best avoided.

By contrast, the central meaning identified in the first paragraph is adverbial, and here the term “objective” is applied to a process of inquiry and/or to an inquirer. This is sometimes referred to as “procedural objectivity,” though this is misleading because it implies mere following of a method. What this adverbial sense of “objectivity” refers to is whether or not inquiry has been **BIASED** through the effects of factors extrinsic to it, notably the researcher’s own motives and interests, social background, political and religious commitments, and so on. It is important to notice that objectivity does not require that these factors have no influence on the research process, only that they do not deflect it from its rational course.

As this makes clear, the adverbial sense of “objective” is evaluative rather than factual in character: It involves judging a course of inquiry, or an inquirer, against some rational standard of how an inquiry *ought to have been pursued in order to maximize the chances of producing true findings*. It is concerned not with all deviations from this path but with systematic error caused by factors associated with the researcher. It is perhaps worth underlining that these factors are not necessarily subjective in the psychological sense.

Objectivity as a regulative ideal, in its adverbial form, has been challenged on a number of grounds. Some commentators have claimed that it can never be achieved, because this would imply a researcher with no background characteristics. On this basis, it has been argued that the notion of objectivity is simply a rhetorical device by which researchers privilege their own views. As already noted, however, objectivity does not require an absence of background characteristics, only that any systematic negative effects of these be avoided. Furthermore, even if it is true that objectivity can never be fully achieved—in that we cannot entirely avoid such negative effects, or at least know that we have avoided them—it does not follow that striving to achieve objectivity is undesirable. It might be argued that the more closely inquiry approximates this ideal, the more likely it is that the knowledge produced will be sound.

It has been proposed that particular categories of actors are more able to gain objectivity than others. One version of this is put forward by Harding (1992). She argues in favor of what she calls “strong objectivity,” which amounts to privileging the perspectives of those who are socially marginalized, on the grounds that they are more capable of gaining genuine knowledge about society because they are not blinded by any

commitment to sustaining the status quo. This approach builds on Hegelian and Marxist ideas to the effect that objectivity is a sociohistorical product, rather than a matter of individual orientation.

Another criticism of objectivity as an ideal stems from rejection of the concepts of rationality and/or truth on which it depends, these being treated as themselves mythological or merely rhetorical. This criticism arises from various forms of skepticism and RELATIVISM, most recently inspired by POSTSTRUCTURALISM.

A rather different criticism of objectivity as a regulative ideal argues that it encourages ethical irresponsibility on the part of researchers. This is because it requires research to be carried out in ways that do not treat important values other than truth as part of its goal. Here, the concepts of truth and rationality are not rejected; the claim is simply that inquiry should involve rational pursuit of other goals besides knowledge, with these other goals—such as social justice—being assumed to have an elective affinity to truth, or as properly overriding it.

Defenders of the older concept of objectivity reject these criticisms, arguing that “objectivity” simply means avoiding bias arising from irrelevant factors, that no one is uniquely privileged or debarred from gaining knowledge of the social world, that any denial of the possibility of knowledge is self-refuting, and that there is good reason to believe that different goals imply divergent lines of rational action and therefore cannot be pursued simultaneously. Defenders of objectivity claim that the success of scientific inquiry has arisen from its institutionalization in research communities, and that seeking to avoid deflection by irrelevant considerations is just as important in research as it is in rational pursuit of any other activity (see Rescher, 1997).

—Martyn Hammersley

REFERENCES

- Harding, S. (1992). After the neutrality ideal: Science, politics, and “strong objectivity.” *Social Research*, 59, 568–587.
- Newell, R. W. (1986). *Objectivity, empiricism and truth*. London: Routledge and Kegan Paul.
- Rescher, N. (1997). *Objectivity: The obligations of impersonal reason*. Notre Dame, IN: University of Notre Dame Press.

OBSERVATION SCHEDULE

Observation schedules are used in research employing observational methods to study behavior and interaction in a structured and systematic fashion (see STRUCTURED OBSERVATION). They consist of pre-defined systems for classifying and recording behavior and interactions as they occur, together with sets of rules and criteria for conducting observation and for allocating events to categories. An observation schedule normally will include a time dimension for locating observations in time and will contain variables relating to different aspects of behavior and interaction, each consisting of a set of categories describing particular behaviors and interactions. The observation schedule is the means by which a researcher operationalizes the aspects of the social world that are the focus of the study; it is used to collect systematic and reliable data that can be used in analysis. The categories and definitions used in observation schedules should be explicit and should be capable of being used identically by different observers.

TYPES OF OBSERVATION SCHEDULE

There are a variety of types of schedule, in particular with regard to how they locate observations in time and whether they provide data on the duration, frequency, time of occurrence, and sequence of the behaviors studied. *Continuous* and *quasi-continuous* observation systems provide a continual, ongoing record of behavior. These systems are the most complete but practical only for limited sets of categories. For example, Hilsum and Cane (1971) studied the work of schoolteachers through continuous observation throughout the day, classifying activities into categories such as “instruction,” “marking,” and “administration.” *Event recording* is essentially a checklist system that notes whenever a particular type of event takes place. It is a good measure of frequency of occurrence. Wheldall and Merrett (1989) used such an approach to compare different ways of managing classroom behavior based on simple counts of every time a teacher praised or reprimanded a pupil. Many systems use various types of time sampling to structure the observations, such as *instantaneous time sampling* and *partial interval* and *whole interval* time sampling. Galton, Simon, and Croll (1980) studied school classrooms using instantaneous

OBLIQUE ROTATION. See ROTATIONS

time sampling; the interactions and behavior of selected pupils and their teacher were coded at precisely timed 25-second intervals. Further details of approaches to time sampling are given by Croll (1986) and Suen and Ary (1989).

DESIGNING OBSERVATION SCHEDULES

Before designing a new observation schedule, researchers should consider whether one of the many systems already in use would serve their purpose. If existing schedules are not suitable, it is still worth considering modifying something already in use or at least incorporating variables from another schedule as benchmark measures for comparative purposes. Guidelines for the development of observation systems are given by Robson (1993, chap. 8). These include criteria such as ease of use in the research situation, low levels of inference or judgment needed by the observer, and explicit definition of categories. The categories used should be *exhaustive*, in that all behaviors encountered can be coded, and *exclusive*, in that each behavior observed fits into only one category of a variable (although there may, of course, be several variables). Researchers intending to use observation schedules may find it helpful to start observing in an unstructured fashion and then to progressively introduce structure into the observation procedure. This can help both to focus their conceptualization of the field of study and to identify ways of more precisely operationalizing the relevant variables.

—Paul Croll

REFERENCES

- Croll, P. (1986). *Systematic classroom interaction*. Lewes, UK: Falmer.
- Galton, M., Simon, B., & Croll, P. (1980). *Inside the primary classroom*. London: Routledge and Kegan Paul.
- Hilsum, S., & Cane, B. (1971). *The teacher's day*. Windsor, UK: National Foundation for Educational Research.
- Robson, C. (1993). *Real world research*. Oxford, UK: Blackwell.
- Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioural observation data*. London: Lawrence Erlbaum.
- Wheldall, K., & Merrett, F. (1989). *Positive teaching: The behavioural approach*. Birmingham, UK: Positive Products.

OBSERVATION, TYPES OF

Observation is a data collection strategy involving the systematic collection and examination of verbal and nonverbal behaviors as they occur in a variety of contexts. This method of data collection is particularly important when there are difficulties in obtaining relevant information through self-report because subjects are unable to communicate (e.g., with infants or confused adults) or provide sufficiently detailed information (e.g., about complex interaction patterns). Observations also are used to validate or extend data obtained using other data collection methods. Both unstructured and structured observations are used by researchers. Unstructured observations are most useful in exploratory, descriptive research. STRUCTURED OBSERVATIONS are used when behaviors of interest are known, and this type of observation often involves the use of an OBSERVATION SCHEDULE (e.g., a checklist). Efforts usually are made to observe participants in as natural a setting as possible. Data collection involving observation involves INFORMED CONSENT of the participants.

There are two main types of observation: observation in person (PARTICIPANT OBSERVATION or NON-PARTICIPANT OBSERVATION) and video recordings. Observation in person involves sustained direct observations by the researcher and focuses on the context as well as the behaviors of individuals to understand the meaning of certain behaviors or beliefs. This type of observation typically is linked with ETHNOGRAPHY. For example, participant observation is a useful method to study privacy in nursing home settings. The participant role or degree of involvement that the researcher has with others in the setting may vary. In nonparticipant observation, researchers focus primarily on the task of observation, while minimizing their participation in interactions in the setting. In other situations, a researcher may be more a participant than observer by taking on some of the roles and responsibilities of insiders in a research setting while observations are carried out. The advantages of in-person observation are that (a) over time, participants accommodate to the presence of the researcher, increasing the likelihood of the possibility of observing the phenomena of interest as it really occurs; (b) the researcher has the opportunity to interact with participants to clarify and extend observations; (c) events can be understood as they unfold in everyday life; and (d) differences between what

participants say and do can be made apparent (Bogdewic, 1999). Observations conducted in this context encompass all the researcher's senses and are recorded as "jottings" in the field, then later are expanded into detailed FIELDNOTES with the researcher's interpretations (Emerson, Fretz, & Shaw, 1995). Usually, it is recognized that participants may limit disclosing some information and define the boundaries of observation.

In the second type of observation, video cameras are used to capture behaviors of interest. The advantage of video recordings is that a permanent record of observations is available for detailed microanalysis or for multiple kinds of analysis (Bottorff, 1994). Video recording typically is used in qualitative ethology and some forms of CONVERSATION ANALYSIS. For example, video recordings provide a rich data source for studying interaction patterns between health care providers and patients. This observational approach is particularly useful when researchers are interested in examining the sequence of behaviors, concurrent behaviors, and nonverbal behaviors that are difficult to observe in real time. In-depth descriptions of behavior can result from the microanalysis of video recorded behaviors.

—Joan L. Bottorff

REFERENCES

- Angrosino, M. V., & Mays de Perez, K. (2000). Rethinking observation: From method to context. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 673–702). Thousand Oaks, CA: Sage.
- Bogdewic, S. P. (1999). Participant observation. In B. J. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (pp. 47–69). Thousand Oaks, CA: Sage.
- Bottorff, J. L. (1994). Using videotaped recordings in qualitative research. In J. M. Morse (Ed.), *Critical issues in qualitative research* (pp. 224–261). Newbury Park, CA: Sage.
- Emerson, R. M., Fretz, R. I., & Shaw, L. L. (1995). *Writing ethnographic fieldnotes*. Chicago: University of Chicago Press.

OBSERVATIONAL RESEARCH

Observation is one of the fundamental techniques of research conducted by sociologists, cultural anthropologists, and social psychologists. It can either stand on its own as a data collection method or be paired

with one or more other means of data collection. The purpose of observational research is to record group activities, conversations, and interactions as they happen and to ascertain the meanings of such events to participants. Observations may take place either in laboratory settings designed by the researcher or in field settings that are the natural loci of selected activities. Data derived from observation may be either encoded quantitatively, often through the use of standardized checklists or research guides, or rendered as open-ended narrative descriptions. Data are recorded in either written or taped (audio and/or video) form and may consist of either numerical or narrative depictions of physical settings, actions, interaction patterns, meanings (expressed or implicit), and expressions of emotion.

Researchers may strive to attempt to accomplish one or more of the following, depending on the type of observation they employ: (a) to see events through the eyes of the people being studied, (b) to attend to seemingly mundane details (i.e., to take nothing for granted), (c) to contextualize observed data in the widest possible social and historical frame (without overgeneralizing from limited data), (d) to attain maximum flexibility of research design (i.e., to use as many means as feasible to collect data to supplement basic observational data), and (e) to construct theories or explanatory frameworks only after careful analysis of objectively recorded data.

TYPES OF OBSERVATIONAL RESEARCH

There are three main ways in which social scientists have conducted observational research. There is considerable overlap in these approaches, but it is useful to distinguish them for purposes of definition: (a) PARTICIPANT OBSERVATION, a style of research favored by field-workers, requires the establishment of rapport in study communities and the long-term immersion of researchers in the everyday life of those communities; (b) reactive observation, associated with controlled settings, assumes that the people being studied are aware of being studied but interact with the researcher only in response to elements in the research design; and (c) unobtrusive (nonreactive) observation is the study of people who are unaware of being observed.

Each of the three types of observational research involves three procedures of increasing levels of specificity:

(a) Descriptive observation involves the annotation and description of all details by an observer

who assumes a nearly childlike stance, eliminating all preconceptions and taking nothing for granted. This procedure yields a huge amount of data, some of which will prove irrelevant. It takes a certain amount of experience in the setting to be able to determine what is truly irrelevant to the participants.

(b) Focused observation entails the researcher looking only at that material pertinent to the issue at hand, often concentrating on well-defined categories of group activity (e.g., religious rituals, political elections).

(c) Selective observation requires honing in on a specific form of a more general category (e.g., initiation rituals; city council elections).

QUESTIONS OF OBJECTIVITY IN OBSERVATIONAL RESEARCH

Social scientists have long realized that their very presence can affect the activities and settings under observation, and they also have endeavored to adhere to careful standards of objective data collection and analysis so as to minimize the effects of observer bias. Objectivity is the result of agreement between participants in and observers of activities and settings as to what is happening; in other words, the researcher has not imposed theoretical or other preconceptions on the events being analyzed.

Social researchers of the first half of the 20th century adopted one of four observational roles of increasing degrees of presumed scientific objectivity: (a) the complete participant, or insider; (b) the participant-as-observer (an insider with scientific training); (c) the observer-as-participant (an outsider who becomes a member of the community, a stance actually preferred by cultural anthropologists); and (d) the complete observer (a researcher without ties to the people or setting being observed who conducts his or her observations unobtrusively, with a minimal amount of interaction with those being observed).

The ideal of complete objectivity, however, has been compromised by modern ethical concerns about INFORMED CONSENT. Researchers who take pains to inform subjects about the nature and purposes of the research, as required by most institutional review boards, would almost certainly be involved in more interaction than the old ideal would accommodate. Modern researchers therefore prefer to characterize their roles in terms of degrees of membership in the setting under observation, on the assumption

that no one can function as an ethical researcher without being a member in some fashion. Membership roles may include (a) peripheral member researchers (those who develop an insider's perspective without participating in activities constituting the core of group membership), (b) active member researchers (those who become involved with the central activities of the group, sometimes even assuming leadership responsibilities, but without necessarily giving a full endorsement of the group's values and goals), and (c) complete member researchers (those who study settings in which they already are members or with which they become fully affiliated as they conduct their research). Researchers in the latter group still, however, strive to make sure that their insider status does not alter the flow of interaction in an unnatural manner—their status as insiders presumably making it possible for them to detect the natural from the unnatural flow of events.

Contemporary observational researchers demonstrate an increased willingness to affirm an existing membership (or to develop a new one) in the group they are studying. They recognize that it might be neither feasible nor desirable to filter out their own perceptions in a quest for the “truth” of a given social situation. In other words, the classic goal of systematic rigor leading to results as nearly objectively rendered as possible has been replaced by an attitude in which multiple versions of “truth” are possible and in which observations are, among other things, records of encounters between specific researchers and specific subjects at specific points in time, rather than reflections of some sort of timeless, abstract social reality. (This attitude shift is more prominent among observational field researchers; experimental researchers, who still have the option of controlling for extraneous variables, continue to strive for rigorous objectivity.)

Field researchers, unlike laboratory researchers, are also redefining the people they study. They prefer to avoid the traditional term for those people, “subjects,” and refer to them instead as “partners,” “collaborators,” or participants in research. The term “subjects” can seem paternalistic, perpetuating an undesirable image of the all-knowing scientist putting human beings under the microscope in a detached, controlled, abstract (and, by implication, unrealistic) manner. Because ethical rules mandate informing “subjects” about their rights and about the nature and scope of a research project, those people are no longer passive actors being observed by a distant authority figure; their

values, beliefs, and ideas must necessarily inform the research.

QUESTIONS OF VALIDITY AND RELIABILITY IN OBSERVATIONAL RESEARCH

Social researchers are becoming increasingly skeptical of using observations of specific people in specific settings to generate large-scale descriptions of cultures supposedly represented by those who are observed. Observational researchers commonly use several techniques to enhance the validity of their data. Cautious about making inferences beyond the study population, yet interested in large-scale patterns, observational researchers may rely on composite portraits generated by standardized observational checklists (see **STRUCTURED OBSERVATION**) used by a team of researchers in many different settings rather than on inferences from a single observed site. It is sometimes suggested that the members of the team should represent a diversity of age, gender, and other demographic characteristics, to make sure that such factors do not inadvertently color any one researcher's perceptions. The "Six Cultures Study," created by John Whiting and his associates in the 1950s to generate cross-culturally valid information about child-rearing practices, is the classic example of this approach. A more recent example is the study of needle hygiene among injection drug users in 20 sites in the United States, led by Stephen Koester, Michael Clatts, and Laurie Price for the National Institute on Drug Abuse.

Data reliability is enhanced by **TRIANGULATION**—the inclusion of observation in a package of data collection methods so as to achieve a reasonably comprehensive conclusion. The fact that one can reach the same conclusion by means of different data collection techniques makes it more likely that the data are reliable.

OBSERVATIONAL RESEARCH IN SELECTED SCHOOLS OF SOCIAL RESEARCH THEORY AND PRACTICE

Observational research has been favored by followers of several major theoretical or conceptual traditions in the social sciences. For example, the school of formal sociology (of which Georg Simmel was an early representative) focuses on the forms or structures that pattern social interactions and relations, rather than on the specific content of those relations. More modern

formal sociologists, such as Manfred Kuhn and Carl Couch, see social reality as constructed by an interaction of the self and the other through more or less institutionalized forms of interconnectedness. Proponents of this school have advocated the use of video technology to record observational data; those with a cultural anthropological orientation favor videotaping in the field, whereas others favor bringing subjects into a specially designed setting in which they can be taped. Such studies tend to focus on the ways in which people develop and sustain relationships, although modifications of the technique also have been used to study the operation of large bureaucracies.

Dramaturgical sociology, associated with the work of Erving Goffman, has focused on how people present themselves in social settings and, in effect, act out their self-conceptions in front of an "audience"—the social group. Goffman's approach, sometimes criticized as unsystematic and hence overly subjective, nonetheless enables the researcher to capture a range of acts of people at all levels of society. Followers of Goffman such as Spencer Cahill have attempted to introduce a degree of systematization in the method; Cahill in particular has made use of teams of researchers, although he has used Goffman's approach of unstructured recording (making notes on the spot and categorizing them afterward, rather than immediately sorting observed behavior into precoded categories) in public and semipublic places. By contrast, Laud Humphreys has attempted to develop a systematic method for capturing dramaturgical detail. In his study of homosexual encounters in a public restroom, he made sure to record his observations on a checklist, although the fact that this method made it possible to identify participants (who, being engaged in a covert activity, probably would not have given permission to be observed so closely) has led critics to express concern about the ethics of such research.

Auto-observation (or **AUTOETHNOGRAPHY**) is a technique that moves the unit of observation from groups in public or semipublic places to the individual and his or her intimate relationships. More often than not, the individual in question is the researcher himself or herself. This approach, sometimes referred to as existential sociology, traces its philosophical roots to sociologist Wilhelm Dilthey, who advocated a stance of *verstehen* (understanding through empathy). An important modern practitioner is Carolyn Ellis, recognized for her powerfully personal account of the death of her partner and the ways in which she dealt with

grief. Although the case example is the sociologist herself, the theme of the study is universal, and the reactions she exhibits are meant to be representative of the culture and society of which she is a part. This approach is linked conceptually to the school of reflexive anthropology, in which participant observers shed their pose of strict objectivity and insert themselves into the narratives of the cultures they are describing. Vincent Crapanzano and Peter Wilson, who structured their ethnographies (of a community in Morocco and of a Caribbean island, respectively) through the lens of their personal relationships with their principal informants, are two prominent practitioners of this approach.

Perhaps no school of social research has placed more emphasis on observational techniques than that of ETHNOMETHODOLOGY. Advocates of this approach believe that social life is constructed in terms of sub-conscious processes—minute gestures that everyone takes for granted and that cannot, therefore, be captured by asking people questions about what they are doing. Ethnomethodologists favor the analysis of minutely detailed data, usually by means of an intricate notational system, an approach they share with the cognitive anthropologists who have turned their attention to the analysis of discourse with an emphasis on conversational overlaps, pauses, and intonations. Sociologist Douglas Maynard used this technique to study the language of courtroom negotiation, and anthropologist Michael Agar used it to study the behavior of drug addicts in a rehabilitation program. Nonverbal discourse also may be recorded by means of this approach, as in the methodology of proxemics (the study of the ways in which people in a given culture use space and how the arrangement of people in space—e.g., the positioning of furniture in a house; the patterns by which people fill in seats in an airport lounge—conveys cultural meanings about relationships), which was pioneered by Edward Hall, and the methodology of kinesics (the academic version of the study that has entered pop psychology as “body language”) developed by R. L. Birdwhistle.

THE ETHICAL DIMENSION OF OBSERVATIONAL RESEARCH

Observational researchers resist the notion that they are engaged in harmfully intrusive work, although, as noted above, they certainly are aware of the necessity of increasing their interactions with members of study populations. Observational research that is

defined as “public” (e.g., studies of patterns in the ways people orient themselves in elevators) may be exempt from institutional review, although examples of purely observational research with minimal interaction are becoming increasingly rare. The more interaction there is, the more opportunities there will be for doing harm in ways that cannot always be adequately anticipated at the outset of a project. It is therefore considered important for observational researchers, regardless of degree of interaction, to be as open as possible with the people they study, and to disclose as much about the project’s methods and aims as they can.

—Michael V. Angrosino

REFERENCES

- Adler, P. A., & Adler, P. (1994). Observational techniques. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 377–392). Thousand Oaks, CA: Sage.
- Agar, M. (1973). *Ripping and running*. New York: Academic Press.
- Angrosino, M. V., & Pérez, K. (2000). Rethinking observation: From method to context. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 73–702). Thousand Oaks, CA: Sage.
- Cahill, S. (1987). Children and civility: Ceremonial deviance and the acquisition of ritual competence. *Social Psychology Quarterly*, 50, 312–321.
- Crapanzano, V. (1980). *Tuhami: Portrait of a Moroccan*. Chicago: University of Chicago Press.
- Ellis, C. (1995). *Final negotiations: A story of love, loss, and chronic illness*. Philadelphia: Temple University Press.
- Gold, R. L. (1958). Roles in sociological field observation. *Social Forces*, 36, 217–223.
- Humphreys, L. (1975). *Tearoom trade: Impersonal sex in public places*. Chicago: Aldine.
- Maynard, D. W. (1984). *Inside plea bargaining: The language of negotiation*. New York: Plenum.
- Price, L. J. (2002). Carrying out a structured observation. In M. V. Angrosino (Ed.), *Doing cultural anthropology* (pp. 107–114). Prospect Heights, IL: Waveland.
- Schensul, J. J., & LeCompte, M. D. (Eds.). (1999). *Ethnographer’s toolkit*. Walnut Creek, CA: AltaMira.
- Whiting, J. W., Child, I. L., & Lambert, W. W. (1966). *Field guide for the study of socialization*. New York: Wiley.
- Wilson, P. J. (1974). *Oscar: An inquiry into the nature of sanity*. New York: Vintage.

OBSERVED FREQUENCIES

A frequency can be defined simply as the number of times an event, which we are measuring, happens.

Table 1 Number of Cars Parked During the Week

<i>Day of the Week</i>	<i>Total Number of Cars Parked</i>
Monday	10
Tuesday	9
Wednesday	10
Thursday	8
Friday	1
Saturday	3
Sunday	2

The event is the VARIABLE on which we are interested in gaining information. Consider the data in Table 1.

The event/variable we are measuring is that of parking lot occupancy. Thus, on Monday, the frequency (i.e., the number of cars parked) is 10, on Tuesday the frequency is 9, on Wednesday it is 10, and so on.

An observed frequency is defined as the *actual* number of times we see the event, which we are measuring, happening. Does this mean we have to stand in the car park and record every car that is parked? In some cases, yes, we would, and could, measure an event by physically being there and counting its frequency. In reality, we tend to gain details about the event/variable we are measuring by using a research instrument (e.g., QUESTIONNAIRE). For example, we could ask the parking lot attendant to fill in a questionnaire about the occupancy of the lot on the days of the week and, using that information, construct Table 1.

We can elaborate on this analysis by looking at the RELATIVE FREQUENCY, which is derived from the observed frequency. The relative frequency is the proportion of cases that include the phenomena we are examining, in relation to the total number of cases in the study.

For example, the total number of cars parked in the parking lot example during the week was 43. We can calculate the relative frequency for each day of the week by using the following formula:

$$\text{relative frequency} = \frac{\text{number of cars parked in the lot on a particular day}}{\text{total number of cars parked during the week}}$$

Thus, on Monday there were 10 cars parked, and the total during the week was 43; therefore, the relative frequency for Monday is as follows:

$$\text{relative frequency (Monday)} = 10/43 = 0.23.$$

From this calculation, we would state that the relative frequency of occupancy on Monday was 0.23.

We could repeat these calculations and compute the relative frequency for the remaining days of the week.

Observed frequency should not be confused with *expected* frequency. Although the two are joined by the fact that they are looking at the same event/variable, they are quite different in their calculation and interpretation.

An expected frequency is *not* an observed frequency, although it may be based upon one. Expected frequencies are what we *expect* to happen, as opposed to what *actually* happens. Expected frequencies usually are based upon some previous information that a researcher has acquired, calculated, or noted, and from which he or she attempts to predict what will happen. For example, we could observe how many cars were parked on four consecutive Mondays and from this attempt to predict what we expect to happen on the fifth Monday.

—Paul E. Pye

REFERENCES

- Kerr, A. W., Hall, H. K., & Kozub, A. (2002). Descriptive statistics. Chapter 2 in *Doing statistics with SPSS*. London: Sage.
- Hinton, P. R. (1996). *Statistics explained: A guide for social science students*. London: Routledge.

OBSERVER BIAS

Although they may strive for OBJECTIVITY in the recording and analysis of data, social researchers who use observational methods are aware of the possibility that bias arising out of the nature of observation itself may compromise their work. It is therefore desirable for researchers to become consciously aware of such possibilities and then to take steps to avoid them.

Observer bias may arise out of unconscious assumptions or preconceptions harbored by the researcher. In some cases, these preconceptions take the form of ethnocentrism—the unreflective acceptance of the values, attitudes, and practices of one's own culture as somehow normative, leading to the inability to see, let alone to understand, behaviors that do not conform to that norm. Feminist scholars have increased researchers' awareness of the unspoken assumptions based in gender and in the power relationships implied by gender roles. Other critics have pointed out ethnocentrism rooted in the social class, religious

background, and even age (the “generation gap”), all of which may contribute to researchers missing important points about the lives of people who are in any way different from themselves.

The effects of ethnocentric observer bias usually can be mitigated by a rigorous and sustained effort at making conscious all the unconscious assumptions that might lead to misperception. Somewhat more subtle—and hence more difficult to overcome—is the bias resulting from the tendency of people to change their behavior when they know they are being observed, even if the researcher doing the observation is doing all the right things to minimize his or her ethnocentrism. Traditional cultural anthropology was based on the establishment of long-term relationships in a study community so that, in effect, the members of the community would forget that they were in fact being observed by someone who had ceased to be a stranger. It is no longer always possible, even for anthropologists, to arrange for long-term field studies. It is therefore necessary to devise other ways to mitigate observer effects in short-term research; for example, the researcher can enhance the level of comfort and trust in the study population by being introduced to the community by someone who already is respected by that group. It is also desirable to practice TRIANGULATION—the collection of impressions gleaned from observation with other sources of data, the better to ensure that unintended cues emanating from any one researcher have not skewed the results.

Most contemporary social researchers recognize that they cannot predict every potential effect that they may have on a study community, and they admit that it is unreasonable to strive to be strictly neutral presences. Some aspects of a researcher’s self-presentation can be easily modified to minimize his or her obtrusiveness in a study community (e.g., style of dress, tone of voice). Other factors (e.g., race, gender, age) cannot be changed; it is therefore ethically desirable for researchers, when reporting results, to be candid about such factors, so that readers can judge for themselves whether or not they seriously compromised the study population’s response to the conditions of research.

—Michael V. Angrosino

REFERENCES

Gouldner, A. W. (1962). Anti-minotaur: The myth of a value-free sociology. *Social Problems*, 9, 199–213.

Schensul, S. L., Schensul, J. J., & LeCompte, M. D. (1999). *Essential ethnographic methods: Observations, interviews, and questionnaires*. Walnut Creek, CA: AltaMira.

ODDS

The odds that a state exists (for example, that someone is a computer user) or that an event occurs (for example, that someone commits a crime) is the ratio of the probability that the state or event occurs to the probability that the state or event does not occur. If we symbolize the probability of the state or event by $P(Y)$, then the odds of Y , $\text{odds}(Y) = P(Y)/[1 - P(Y)]$. If the event or state is impossible, the probability is zero and the odds are zero. If the probability is less than 0.5, the odds will be between zero and 1; if the probability is exactly 0.5, the odds will be 1; if the probability is greater than 0.5, the odds will be greater than 1; and if the state or event is certain, $P(Y) = 1$ and $\text{odds}(Y) = 1/0$, which is positive and infinitely large in magnitude. The idea of the odds of an outcome, or the comparison of the odds of two alternative outcomes, is common in gambling, which provided some of the impetus for the early development of probability theory and statistics (Stigler, 1986, pp. 62–63).

The odds is the ratio of two “opposite” probabilities, the first equal to one minus the second. Another related ratio of two probabilities is the *relative risk* or *risk ratio*, the ratio of two different but not necessarily opposite probabilities. For example, if the probability of computer use is .60 for males, then the odds of computer use for males, $\text{odds}(Y)$, is $(0.60)/(1 - 0.60) = 1.5$; if the probability of computer use is .45 for females, then the odds of computer use for females is $(0.45)/(1 - 0.45) = 0.82$, and the ratio of the probabilities (or risks) of computer use for males compared to females is the relative risk $= 0.60/0.45 = 1.33$. In other words, males are 1.33 times as likely to use computers as females. A second related ratio, the ODDS RATIO, is the ratio of two different but not opposite odds. In the above example, the odds ratio of males as opposed to females being computer users is $\text{odds ratio} = (\text{odds of male using computer})/(\text{odds of female using computer}) = 1.5/0.82 = 1.83$. In other words, the odds of computer use is 1.83 times as high, or 83% higher, for males as for females. Notice that the relative risk and the odds ratio are not the same. A common mistake is to calculate an odds ratio, then interpret it

as though it were a relative risk, for instance, in the above example, saying that males are 1.83 times as likely to use computers as females (when, as shown above, males are only 1.33 times as likely as females to be computer users). It is important to keep clear the distinction between (a) probabilities or ratios of probabilities and (b) odds or ratios of odds.

—Scott Menard

REFERENCES

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press and Harvard University Press.

ODDS RATIO

An odds ratio is the division of two ODDS. An odds is formed by two numbers representing two different events (most often event vs. nonevent), such as the odds of winning versus losing, or the odds of voting for the Democratic candidate versus voting for the Republican candidate (or a non-Democratic candidate). As such, it is different from a rate, which gives the number of events occurring in a population standardized to a base, such as per hundred or per thousand. Odds ratios are useful statistical tools in CATEGORICAL DATA ANALYSIS for studying association or concordance in the data.

G. Udny Yule was notably the first statistician to apply odds ratios to measure the association in CONTINGENCY TABLES. Odds ratios and log-odds ratios are common tools for LOG-LINEAR MODELING. To demonstrate the application of odds ratios, let us examine in Table 1 some classic data presented by M. Greenwood and G. U. Yule in their classic 1915 study of the effect of antityphoid inoculations.

This is a typical 2×2 table with the columns recording the two outcomes and the rows representing the two categories in the exposure variable. More generally, the columns may contain any response or dependent variable and the rows, any explanatory or

Table 1 Data on Antityphoid Inoculations

	<i>Typhoid: No</i>	<i>Typhoid: Yes</i>	<i>Total</i>
Inoculation: Yes	6,759	56	6,815
Inoculation: No	11,396	272	11,668
Total	18,155	328	18,483

SOURCE: Data are from Greenwood and Yule (1915).

Table 2 A Two-by-Two Table Layout

	<i>Variable 1, Category 1</i>	<i>Variable 1, Category 2</i>	<i>Total</i>
Variable 2, Category 1	<i>A</i>	<i>B</i>	<i>A + B</i>
Variable 2, Category 2	<i>C</i>	<i>D</i>	<i>C + D</i>
Total	<i>A + C</i>	<i>B + D</i>	<i>A + B + C + D</i>

independent variable; indeed, the rows and columns can simply contain two variables whose association is under investigation. The cells can be indicated by *A*, *B*, *C*, and *D* (see Table 2).

There are three nonredundant ways of forming odds ratios (OR) in a 2×2 table:

$$OR_1 = \frac{A/C}{D/B}, \quad OR_2 = \frac{A/B}{D/C}, \quad \text{and} \quad OR_3 = \frac{A/B}{C/D}.$$

For studying association between two categorical variables, OR_3 typically is used. This odds ratio is also known as the cross-product ratio because

$$OR = \frac{A/B}{C/D} = \frac{AD}{BC},$$

where hereafter we ignore the subscript because it is the only odds ratio we examine. Returning to the data on antityphoid inoculations, we find the odds ratio or cross-product ratio to be

$$OR = \frac{(6,759)(272)}{(56)(11,396)} = 2.881.$$

The ratio suggests that there is an association between inoculations and typhoid occurrences. The odds of having typhoid prevented are close to three times greater for those who had received inoculations than for those who had not. Statistics like these may give some evidence to government agencies for policy making such as implementing vaccination programs.

Note that YULE'S *Q* is simply a function of the odds ratio that has been normed to fall between -1 and $+1$:

$$Q = \frac{OR - 1}{OR + 1} = \frac{2.881 - 1}{2.881 + 1} = 0.485.$$

Because the *Q* statistic is close to 0.5 (with zero indicating no relationship), it suggests some moderate positive association between inoculations and typhoid prevention.

Often, it is convenient to work with (natural-) log-odds ratios, which have two useful properties. Log-odds ratios have zero as the value for indicating no

relation between the two variables, whereas the value for odds ratios is one. Zero for no relationship may be intuitively appealing for some researchers. One also can compute asymptotic standard errors for log-odds ratios,

$$\begin{aligned}\sigma_{\log(\text{OR})} &= \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}} \\ &= \sqrt{\frac{1}{6,759} + \frac{1}{56} + \frac{1}{11,396} + \frac{1}{272}} = 0.148,\end{aligned}$$

and calculate related Z ratios that follow the standard normal distribution:

$$Z = \frac{\log(\text{OR})}{\sigma_{\log(\text{OR})}} = \frac{0.460}{0.148} = 3.108.$$

The Z statistic indicates a highly significant result (i.e., it is very unlikely that the odds is observed by chance). One can go one step further by constructing confidence intervals for log-odds ratios using the Z obtained (or for odds ratios by going through the anti-log function).

—Tim Futing Liao

REFERENCE

- Greenwood, M. & Yule, G. U. (1915). The statistics of antityphoid and anticholera inoculations, and the interpretation of such statistics in general. *Proceedings of the Royal Society of Medicine* 8 (part 2), 113–194.
- Rudas, T. (1997). *Odds ratios in the analysis of contingency tables*. Thousand Oaks, CA: Sage.

OFFICIAL STATISTICS

Counts made by administrators on behalf of their rulers for the purposes of gathering taxes or conscripting soldiers were a feature of many ancient civilizations, but the organized compilation of facts and figures for the use of states in much broader ways became systematized only following the age of Enlightenment, the scientific revolution, and the growth of European nationalism. Nowadays, modern states collect and publish a whole range of DATA about their populations, trade, finances, housing, health, consumption, leisure activities, and many other aspects of their societies, both to reflect performance and as a guide for future policy formation. In addition, supranational bodies

such as the European Commission, the United Nations and its subsidiary organizations, the World Health Organization, and the World Bank all monitor their areas of competence, to produce a broader picture of regional and global developments.

MAIN TYPES OF OFFICIAL STATISTICS

Some official statistics are generated routinely as a result of administrative procedures, and there often is a legal requirement for information to be provided to the state. Tax details, births, marriages, divorces, deaths, and certain notifiable diseases are recorded in this way, as are data about employment and unemployment, crime, and education. STATISTICS about these collections of data are reported at regular intervals in many countries.

Although apparently comprising a full record, such statistics have often been questioned on the grounds that they provide an incomplete picture of the areas they claim to cover. Most vulnerable to this type of criticism have been official figures about crime, employment, and unemployment. Alternative means of gathering information about crime, such as victim SURVEYS, have revealed the gulf between levels of crime experienced by the population and published statistics of recorded crime. This discrepancy, known as the “dark figure,” stems partly from the fact that many offenses are not reported, for a wide variety of reasons including disbelief by victims or witnesses that effective action will be taken, fear of retaliation by perpetrators, or simply apathy (Coleman & Moynihan, 1996).

Similarly, official employment statistics fail to take account of those working in the informal or illegal economy, and unemployment statistics omit those who have become discouraged about finding work and have dropped out of the labor market, as in depressed market situations when jobs are scarce. Because unemployment rates are politically sensitive, governments may take administrative measures to reduce still further the reported numbers out of work, such as by excluding the long-term unemployed and others not eligible for compensatory benefits.

The most widely known and longest established form of official statistics are the results of the CENSUS of population, a compendium of information about all inhabitants of a state. The U.S. Bureau of the Census conducts the national census at 10-year intervals. Although a census is regarded as the most comprehensive source of information about the entire population

of a country, the broad scope of its focus usually renders it inadequate for more detailed investigation.

Recent decades have seen a growth of government surveys, usually targeted at specific topic areas to investigate inevitable gaps left by censuses. For reasons of cost, if no other, many kinds of official statistics generated in this way are not based on all of those from whom data might be gathered. Instead a subset, consisting of a large-scale REPRESENTATIVE SAMPLE, is approached, and consequent findings are then generalized to provide estimates about the whole POPULATION. Using either INTERVIEWING or mail QUESTIONNAIRES to gather data, these surveys often achieve high response rates, inspiring confidence in their inferences.

Typical examples of such surveys in the United Kingdom are the Family Expenditure Survey, established as early as the 1950s, and the 1970s General Household Survey. As its name implies, the latter is a general purpose data-gathering tool for use by a wide range of researchers and is flexible enough to be modified by the introduction of new questions to take account of changing social trends or policy priorities. Other major surveys include cohort studies such as the Longitudinal Survey, which tracks a 1% sample of the U.K. population, the Labour Force Survey, and the British Household Panel Survey.

These last two studies are examples of state-sponsored surveys designed to be compatible with those of other countries so that results contribute to international statistical reports. Such cross-national statistics are increasingly the norm, and the adoption of standard measures provides not only comparability but also safeguards against manipulation by individual governments. Nevertheless, problems remain, and where statistics relate to politically sensitive issues, such as asylum seekers, this can affect their dissemination (Singleton, 1999).

CHANGING ACCESS TO OFFICIAL STATISTICS

During the last decades of the 20th century, the significance of the long-standing practice of gathering and compiling official statistics was completely transformed. Although the census of population had long been automated by the use of punched card processing (1880 in the United States and 1911 in the United Kingdom), in the later years of the 20th century the expansion in computing capacity available to the ordinary citizen through the rapid spread and quantum

leap in processing power of personal computers was matched by the explosive growth of the World Wide Web and the proliferation of databases holding vast amounts of statistical information. These parallel processes meant that it was no longer necessary to rely entirely on the published accounts of governmental or other official agencies: Individuals now had the potential to interrogate data sets for themselves and make their own interpretations.

These technological developments did not take place in a political vacuum. Governments, particularly in OECD countries, embraced a democratizing agenda to make more information available to their citizens so that they could gain a fuller understanding of the changing nature of their society and participate more actively in debates about the governance of their states. Many governments and organizations now provide free access to databases held on their Web sites. Meanwhile, international statistics are coordinated and made available by supranational agencies such as Eurostat. At the same time, it should be recognized that ways in which data are produced and released place subtle constraints on this new paradise where a wealth of apparently reliable and objective information is freely available.

THE POLITICAL CONTEXT OF OFFICIAL STATISTICS

Attempts by the U.K. government in the 1980s to suppress the publication of research findings that revealed a growing mortality gap between rich and poor, as well as problems such as those about unemployment statistics discussed above, serve as reminders of the highly sensitive political context of some statistics (Townsend, Davidson, & Whitehead, 1988). It should never be forgotten that although certain published figures are termed "official," this offers no guarantee of their accuracy or impartiality or that hidden censorship or pressure has not occurred. The issue of impartiality remains a concern for all users of official statistics worldwide, and in the United Kingdom the debate continues despite government pledges to freedom of information and an independent national statistical service.

—Will Guy

REFERENCES

- Coleman, C., & Moynihan, J. (1996). *Understanding crime data: Haunted by the dark figure*. Buckingham, UK: Open University Press.

Singleton, A. (1999). Measuring international migration: A case study of European cross-national comparisons. In D. Dorling & S. Simpson (Eds.), *Statistics in society: The arithmetic of politics* (pp. 148–158). London: Arnold.

Townsend, P., Davidson, N., & Whitehead, M. (1988). *Inequalities in health: The Black report and the health divide*. Harmondsworth: Penguin.

OGIVE. See CUMULATIVE FREQUENCY POLYGON

OLS. See ORDINARY LEAST SQUARES

OMEGA SQUARED

Omega squared (ω^2) is a measure of the strength (or magnitude) of an effect. There are various measures of the strength of an effect relevant to different data measurements (e.g., the CORRELATION COEFFICIENT is one kind of effect strength measure). Omega squared is relevant to experimental studies that have been analyzed using one-way or factorial ANALYSIS OF VARIANCE. It is one of a group of measures of the strength of effect for analysis of variance that are related to *R*-SQUARED (R^2) (the measure of the strength of effect in regression and correlation). The strength of an effect is scientifically important because it gives a comparative measurement of how much a treatment (in an experimental study) has affected the behavioral measure. An effect size of zero means the treatment had no effect, and measures of strength of effect related to *R*-squared have a maximum of 1 (as does the correlation coefficient).

In a one-way (unrelated measures) analysis of variance, the results table looks like Table 1.

Table 1 A One-Way Analysis of Variance

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio
Treatment	k	a	x	$\frac{x}{y}$
Error	$n - k - 1$	b	y	
Total	$n - 1$	c		

NOTE: k is the number of treatment groups and n is the total number of measurements (in social science research, usually participants).

Table 2 Example Results From a One-Way Analysis of Variance

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio
Treatment	3	600	200	2
Error	20	2,000	100	
Total	23	2,600		

For fixed treatment effects, the mathematical definition of ω^2 is

$$\omega^2 = \frac{a - (k - 1)y}{c + y}.$$

Suppose a one-way ANOVA produced results like Table 2.

Then

$$\omega^2 = \frac{600 - (3 - 1)100}{2,600 + 100}.$$

The definition for random effects differs because the variance component calculation is different from that for fixed effects.

A more general formula is

$$\omega_{\text{effect}}^2 = \frac{\hat{\sigma}_{\text{effect}}^2}{\hat{\sigma}_{\text{total}}^2}.$$

where $\hat{\sigma}_{\text{effect}}^2$ is the VARIANCE COMPONENT for an effect. This formula can be used for fixed and random effects, assuming the variance components are calculated appropriately (e.g., see Howell, 2002).

There are other potential measures of effect strength for analysis of variance. These include ETA squared (η^2), which is also related to R^2 , and phi prime (ϕ'), which is a measure of standardized effect size, related to Cohen's d .

Omega squared is a less biased measure of magnitude of effect than is eta squared (see Fowler, 1985). This means it is a better estimate of the treatment effect in the population, as opposed to the magnitude of the effect found in the current sample. Insofar as there are differences between ω^2 and η^2 , ω^2 is always smaller and more useful when we are interested in generalizing results to a population.

—Martin Le Voi

See also BIAS, ETA, FIXED-EFFECTS MODEL, *R*-SQUARED, VARIANCE

REFERENCES

- Fowler, R. L. (1985). Point estimates and confidence intervals in measures of association. *Psychological Bulletin*, 98, 160–165.
- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Duxbury, UK: Thomson Learning.

OMITTED VARIABLE

In any research situation, the estimated effect of one variable on another may change when a third variable is introduced. The direction of a relationship between two variables may change or the relationship may disappear altogether when controlling for a third variable. Social scientists must beware the danger of excluding relevant variables from a research design. Failure to account for relevant variables may result in omitted variable bias.

Suppose we want to estimate the effect of INDEPENDENT VARIABLE X_1 on DEPENDENT VARIABLE Y . The effect of X_1 on Y is denoted as β_1 . If we use REGRESSION analysis, our estimate of β_1 is b_1 . If omitted variable bias is present in our research, our estimate b_1 is incorrect because we have failed to account for another independent variable, X_2 .

How do we know if omitting X_2 biases our estimate of β_1 ? If X_2 has no effect on the dependent variable Y , then omitting X_2 will not induce bias. In other words, if an independent variable is irrelevant to the phenomenon under study, then excluding that independent variable will not cause bias.

An omitted variable will bias our estimate of β_1 only if the omitted variable is correlated with our explanatory variable X_1 . Introducing X_2 into the regression equation will change our estimate of β_1 only if X_2 is correlated with X_1 . If X_2 is uncorrelated with X_1 , we can safely omit it from our analysis, even if X_2 is strongly related to Y . Thus, we can omit variables that are not correlated with our explanatory variables, even if the omitted variables have a strong relationship with the dependent variable. We will induce omitted variable bias only if we exclude variables that are correlated with the included explanatory variables.

However, if our goal is predicting values of Y , then we should include in our regression every explanatory variable that is strongly related to the dependent variable. Explaining as much variation in the dependent variable as possible will increase the accuracy of

our predictions, and we will explain more variation if we include all relevant explanatory variables. When FORECASTING is our goal, we should include all independent variables strongly related to Y . If we do not want to forecast values of Y but simply want to know the effect of X_1 on Y , we can safely exclude any variables not correlated with X_1 .

—Megan L. Shannon

REFERENCE

- King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

ONE-SIDED TEST. See ONE-TAILED TEST

ONE-TAILED TEST

A NULL HYPOTHESIS most often specifies one particular value of a population PARAMETER. We do not necessarily believe this to be the true value of the parameter, but if we can reject the null hypothesis and thereby eliminate this particular value of the parameter, then it follows that the parameter must be equal to something else. In the study of the relationship between two variables, using REGRESSION as an example, the null hypothesis is most often stated in such a way that there is no relationship between the variables. In symbols, $H_0: \beta = 0$, where β is the SLOPE of the population regression line.

If we can reject this null hypothesis, then we have shown that β is not equal to 0, meaning that a relationship exists between the variables. Next, if the population slope is not equal to 0, then what can we conclude about the value of this slope? One possibility is that we do not know anything about the slope, meaning that it can be either greater than or less than zero. Another possibility is that we have additional knowledge about the slope. Say that we know, from previous research, that the slope cannot be negative. The ALTERNATIVE HYPOTHESIS can then be stated as $H_1: \beta > 0$. Because the null hypothesis has been rejected, the conclusion follows from the alternative hypothesis that the value of β must be larger than 0.

This is an example of a *one-tailed* (also called one-sided) test. It is called one-tailed, or one-sided, because of the one-sided form of the alternative hypothesis. The alternative hypothesis includes values only in one direction away from the value of the parameter of the null hypothesis.

When there is a one-tailed alternative hypothesis, the null hypothesis gets rejected for only one range of the test statistic. When we have a normal test statistic and a 5% SIGNIFICANCE LEVEL, a null hypothesis with a two-tailed alternative hypothesis gets rejected for $z < -1.96$ or $z > 1.96$. Half of the significance level is located at each tail of the test statistic, and we reject for large, negative values or large, positive values. For a one-tailed test with a 5% significance level, however, the null hypothesis is rejected for $z > 1.64$. That means the rejection region is located here only in one (the positive) tail of the distribution of the test statistic. This is because we make use of the additional knowledge we have about the parameter—that it is greater than zero.

The distinction between a two-tailed and a one-tailed alternative hypothesis could matter for a normal test statistic, say if $z = 1.85$. With a two-tailed alternative here, the null would not be rejected, but it would be rejected for a one-tailed alternative. This situation, however, does not occur very often.

Also, with the change from a prechosen significance level to a *P* VALUE computed from the data, the distinction is not as important. If we report a one-tailed *p* value, then the reader can easily change it to a two-tailed *p* value by multiplying by 2. Statistical software must be clear on whether it computes one-tailed or two-tailed *p* values.

—Gudmund R. Iversen

REFERENCES

- Mohr, L. B. (1990). *Understanding significance testing* (Sage University Papers series on Quantitative Applications in the Social Sciences, series 07-073). Newbury Park, CA: Sage.
- Morrison, D. E., & Henkel, R. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.

ONE-WAY ANOVA

Analysis of variance, which is abbreviated as ANOVA, was developed by Sir Ronald Fisher (1925)

to determine whether the means of one or more factors and their interactions differed significantly from that expected by chance. “One-way” refers to an analysis that involves only one factor. “Two-way” (or more ways) refers to an analysis with two or more factors and the interactions between those factors. A factor comprises two or more groups or levels. Analysis of variance is based on the assumption that each group is drawn from a population of scores that is normally distributed. In addition, the variance of scores in the different groups should be equal, or homogeneous.

The scores in the different groups may be related or unrelated. *Related* scores may come from the same cases or cases that have been matched to be similar in certain respects. An example of related scores is where the same variable (such as job satisfaction) is measured at two or more points in time. *Unrelated* or *independent* scores come from different cases. An example of unrelated scores is where the same variable is measured in different groups of people (such as people in different occupations). An analysis of variance consisting of only two groups is equivalent to a *t*-test where the variances are assumed to be equal.

A one-way analysis of variance for unrelated scores compares the population estimate of the variance between the groups with the population estimate of the variance within the groups. This ratio is known as the *F*-test, *F*-statistic, or *F*-value in honor of Fisher and is calculated as

$$F = \frac{\text{between-groups variance estimate}}{\text{within-groups variance estimate}}.$$

The bigger the between-groups variance is in relation to the within-groups variance estimate, the larger *F* will be and the more likely it is to be statistically significant. The statistical significance of *F* is determined by two sets of degrees of freedom: those for the between-groups variance estimate (the numerator or upper part of the *F* ratio) and those for the within-groups variance estimate (the denominator or lower part of the *F* ratio). The between-groups DEGREES OF FREEDOM are the number of groups minus 1. The within-groups degrees of freedom are the number of cases minus the number of groups.

In a one-way analysis of variance for related scores, the denominator of the *F* ratio is the population estimate of the within-groups variance minus the population estimate of the variance due to the cases. The degrees of freedom for this denominator are the number

of cases minus 1 multiplied by the number of groups minus 1.

Where the factor consists of more than two groups and F is statistically significant, which of the means differ significantly from each other needs to be determined by comparing them two at a time. Where there are good grounds for predicting the direction of these differences, these comparisons should be carried out with the appropriate t -test. Otherwise, these comparisons should be made with a post hoc test. There are a number of such tests, such as the SCHEFFÉ TEST.

—Duncan Cramer

REFERENCE

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver and Boyd.

ONLINE RESEARCH METHODS

The development of computer technology and the Internet has greatly enhanced the ability of social scientists to find information, collaborate with colleagues, and collect data through computer-assisted experiments and surveys. A number of different tools and techniques fall within the rubric of online research methods, including Internet search engines, online databases and data archiving, and online experiments. Although the Internet itself, as a collection of networked computer systems, has existed in some form for at least 25 years, the development of the World Wide Web in the early 1990s significantly improved its usability and began the explosion of interconnected information. At the same time, the development of powerful personal computers has brought to the desktop the ability to develop sophisticated data analyses. Combined, these technological advances have provided social scientists with an important array of new tools.

INTERNET SEARCH ENGINES

The Internet is essentially a collection of interconnected computer networks, which themselves are made up of any number of individual computer systems. Information on these systems is made available to Internet users by placing it on a server, of which there are several types, including Web, news, and file transfer protocol (FTP) servers. As of 2002, there

were approximately 200,000,000 servers (or hosts) accessible through the Internet. In order to find material within this overwhelming collection, a wide range of search engines have been developed. A search engine takes the keywords or a sentence entered and returns a list of Web pages that contain the requested words. Some search engines, such as Google (www.google.com) and AltaVista (www.altavista.com) are primarily that; they incorporate a very simple user interface and attempt to return matches that are best using some type of best-fit algorithm. Other engines, such as the one associated with Yahoo! (www.yahoo.com) are more comprehensive, combining an attempt to organize Web sites according to categories, as well as keyword and category searching. Other well-known search engines include Lycos (www.lycos.com) and AskJeeves (www.askjeeves.com). Like Yahoo!, Lycos aims to be more than a search engine, providing a portal service, where users can not only search the Web but also gain access to news and a range of online services. Recent enhancements in search engine technology include options to search for graphical materials such as pictures, as well as for text in word processing documents embedded within Web sites.

Using Internet search engines is as much an art as a science, because for many searches, hundreds and even thousands of potential matches may be returned. In general, the more specific the phrase or keywords used, the better, but the state of search engine technology is such that users often have to wade through many matches to find exactly what they need. In fact, an entire Internet site, www.searchenginewatch.com, is dedicated to helping researchers get the most out of the many different search engines.

ONLINE DATABASES/DATA ARCHIVING

The development of the Internet has greatly improved the ability of scientists to share data. Many data sets now reside on the World Wide Web and are accessible through generalized or specialized search engines. The National Election Studies, for example, provide free access to the full set of data collected since 1948 (www.umich.edu/~nes). Researchers need only register with the site to gain access. The General Printing Office (GPO) of the U.S. federal government has created a single access point to search a range of online databases, including the *Congressional Record* and the *Federal*

Register (www.access.gpo.gov/su_docs/multidb.html). This site also points up a typical problem with online databases—their lack of completeness. The GPO *Federal Register* database, for example, does not begin until 1994, and some of the other constituent data sets are even more recent. Thus, researchers wishing to access significant historical material often cannot find it online. Even so, data sets as broad based as the U.S. Census (www.census.gov) and as specific as the Victorian Database Online (www.victoriandatabase.com) have radically changed the way researchers find information.

For academic researchers, many journals have begun online full-text databases of their articles, maintained either specifically by the journal or as part of a collection of journals. JSTOR (www.jstor.org) is one large full-text database, incorporating page images of 275 journals incorporating more than 73,000 issues, dating back to the first issue for many journals. An even more comprehensive service is provided by EBSCO (www.ebscohost.com), providing access to full text for some 1,800 periodicals and the indices for 1,200 more. Whereas JSTOR provides access back to Volume 1 for many journals, EBSCO full text dates only to 1985 but makes up for this by providing access to current issues, whereas JSTOR limits availability to issues at least 3 years old. Most academic libraries provide users with the ability to use these services, as well as a wide range of other services, including Lexis-Nexus (legal research, news periodicals, and reference information) and ERIC (education).

ONLINE EXPERIMENTS

Although experimentation has been a standard tool in some social science fields for many years, the ability to mount experiments online increases the number and variety of subjects who can participate as well as the sophistication that can be used in the design of the experiment. Computers can be used to manipulate textual and nontextual materials presented to subjects while tracking the way in which the subjects respond. Using computers to mount an experiment simplifies data collection because the experiment can be designed to directly generate the data set as the experiment proceeds, thus avoiding data entry errors. Researchers in decision making and in experimental economics have pioneered much of the use of online experiments, perhaps because their fields lend themselves to simpler experimental designs and experiments that

can be completed quickly. Even so, it is likely that the number of INTERNET SURVEYS is far greater than the number of true experiments running on the Internet. A Hanover College psychology department site (<http://psych.hanover.edu/research/exponnet.html>) maintains a listing of known Internet-based psychology projects, many of which are surveys rather than experiments.

Even so, the use of the Internet for experimentation is growing rapidly. Software such as MediaLab (www.empirisoft.com/medialab/) and Web-based guides (Kevin O'Neil, *A Guide to Running Surveys and Experiments on the World-Wide Web*, <http://psych.unl.edu/psychlaw/guide/guide.asp>) can simplify the process of generating online experiments. Researchers who wish to use this technique must consider several points, however. Will the experiment lend itself to being run without direct personal guidance from the experimenter? Participants will have to negotiate their own way through the experiment using only whatever information can be provided on the Web site. Can the experiment be completed reasonably quickly? Most existing online experiments are relatively short; Web users often will not devote significant time to participation in such experiments. Can the experiment be designed to ensure random assignment to experimental conditions? Although RANDOM ASSIGNMENT is important in any EXPERIMENT, the recruitment of subjects in Internet experiments cannot be controlled as easily as in traditional research, so random assignment is potentially critical. An especially interesting challenge comes in acquiring INFORMED CONSENT from potential subjects. For experiments carried out with students for instructional purposes, this is not an issue, but for data collection to be used in research projects aimed at publication and using the general public as subjects, informed consent must be obtained. The standard approach has become a front end to the experiment that provides the consent document and requires subjects to click on a button on the screen to indicate assent to the document. PsychExperiments at the University of Mississippi (<http://psychexps.olemiss.edu/AboutPE/IRB%20procedures1.htm>) uses this technique, along with the requirement that subjects affirmatively click on a Send Data button to have their results included in the data set.

—David P. Redlawsk

See also COMPUTER-ASSISTED PERSONAL INTERVIEWING, INTERNET SURVEYS

REFERENCES

- Birnbaum, M. H. (2001). *Introduction to behavioral research on the Internet*. Upper Saddle River, NJ: Prentice Hall.
- Krantz, J. H., & Dalal, R. (2000). Validity of Web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–60). New York: Academic Press.
- Musch, J., & Reips, U. (2000). A brief history of Web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 61–88). San Diego, CA: Academic Press.
- Nesbary, D. K. (2000). *Survey research and the World Wide Web*. Boston, MA: Allyn and Bacon.
- University of Mississippi. (n.d.). PsychExperiments. Retrieved from <http://psychexps.olemiss.edu>

ONTOLOGY, ONTOLOGICAL

Ontology is a branch of philosophy that is concerned with the nature of what exists. It is the study of theories of being, theories about what makes up reality. In the context of social science: All theories and methodological positions make assumptions (either implicit or explicit) about what kinds of things do or can exist, the conditions of their existence, and the way they are related.

Theories about the nature of social reality fall into two categories: What exists is viewed either as a set of material phenomena or as a set of ideas that human beings have about their world. The materialist position assumes that, from a human viewpoint, both natural and social phenomena have an independent existence, and both types of phenomena have the potential to constrain human actions (see REALISM). Natural constraints may be in the form of gravity, climate, and our physical bodies, whereas social constraints include culture, social organization, and the productive system. Materialism is associated with the doctrine of naturalism, which claims that because there is little difference between the behavior of inanimate objects and that of human beings, the logics of enquiry appropriate in the natural sciences can also be used in the social sciences. The idealist position, on the other hand, claims that there are fundamental differences between natural and social phenomena, that humans have culture and live in a world of their shared interpretations

(see IDEALISM). Social action is not mere behavior but instead involves a process of meaning giving. It is these meanings that constitute social reality (Johnson, Dandeker, & Ashworth, 1984, pp. 13–15).

Ontological assumptions underpin all social theories and methodological positions; they make different claims about what exists in their domain of interest. For example, POSITIVISM and FALSIFICATIONISM entail ontological assumptions of an ordered universe made up of discrete and observable events. Human activity is regarded as observable behavior taking place in observable, material circumstances. Social reality is regarded as a complex of causal relations between events that are depicted as a patchwork of relations between variables. INTERPRETIVISM, on the other hand, assumes that social reality is the product of processes by which human beings together negotiate the meanings of actions and situations. Human experience is characterized as a process of interpretation rather than direct perception of an external physical world. Hence, social reality is not some “thing” that may be interpreted in different ways; it is those interpretations (Blaikie, 1993, pp. 94, 96).

Because ontological claims are inevitably linked with epistemological claims (see EPISTEMOLOGY), it is difficult to discuss them separately (Crotty, 1998, p. 10). Assertions about what constitutes social phenomena have implications for the way in which it is possible to gain knowledge of such phenomena. Differences in the types of ontological and epistemological claims cannot be settled by empirical enquiry. Although they are open to philosophical debate, the proponents of the various positions ultimately make their claims as an act of faith.

—Norman Blaikie

REFERENCES

- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity Press.
- Crotty, M. (1998). *The foundations of social research*. London: Sage.
- Johnson, T., Dandeker, C., & Ashworth, C. (1984). *The structure of social theory*. London: Macmillan.

OPEN QUESTION. See OPEN-ENDED QUESTION

OPEN-ENDED QUESTION

Open-ended questions, also called open, unstructured, or qualitative questions, refer to those questions for which the response patterns or answer categories are provided by the respondent, not the interviewer. This is in contrast to closed-ended or structured questions, for which the interviewer provides a limited number of response categories from which the respondent makes a selection. Thus, respondents can provide answers to open questions in their own terms or in a manner that reflects the respondents' own perceptions rather than those of the researcher.

This type of question works well in face-to-face interviews but less well in mail, electronic, or telephone interviews. The open-ended question is effective particularly when phenomenological purposes guide the research. People usually like to give an opinion, making the open-ended question potentially therapeutic for the respondent. In addition, open-ended questions often will generate unanticipated accounts or response categories. The open-ended question also is appropriate when the range of possible responses exceeds what a researcher would provide on response list (e.g., recreation preference); when the question requires a detailed, expanded response; when the researcher wants to find out what a respondent knows about a topic; and when it is necessary to flesh out details of a complicated issue (Fowler, 1995). Open-ended questions also are useful in generating answer categories for closed or structured questions.

The open-ended question is, however, more demanding on respondents, particularly those with less education (Frey, 1989). The detail and depth of a response are also affected by the extent to which a respondent is familiar with or has an investment in the research topic. An open-ended question can prompt a lengthy, detailed response, much of which might be irrelevant to the topic, and which may be difficult to code. Open-ended questions in interviews also typically take longer to ask and to record responses, thereby increasing survey costs. Cost is perhaps the primary reason that the open-ended question is utilized in a limited way in survey research today; survey researchers will use this type of question when it is the only way certain responses can be obtained.

Finally, the open-ended question becomes a "linguistic event" where the interviewer is not

simply a neutral or passive conduit of the research questions. Rather, the interview becomes a negotiated text reflecting the interaction of respondent and interviewer (Fontana & Frey, 2000). The interview is a social encounter amplified by the open-ended questions, and the process has the potential to provide an outcome that is not representative of a respondent's perception of events; it is, in a sense, distorted by the interaction of interviewer and respondent.

Open-ended questions are valuable means of obtaining information despite problems associated with burden, cost, and context effects. Some of the difficulties can be overcome by adding some structure or limits to responses (Fowler, 1995) and through proper interviewer training on the administration of open-ended questions.

—James H. Frey

REFERENCES

- Fontana, A., & Frey, J. H. (2000). The interview: From structured questions to negotiated text. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 645–672). Thousand Oaks, CA: Sage.
- Fowler, F. J. (1995). *Improving survey questions*. Thousand Oaks, CA: Sage.
- Frey, J. F. (1989). *Survey research by telephone*. Newbury Park, CA: Sage.

OPERATIONAL DEFINITION. See
CONCEPTUALIZATION, OPERATIONALIZATION,
AND MEASUREMENT

OPERATIONALISM/OPERATIONALISM

Operationalism began life in the natural sciences in the work of Percy Bridgman and is a variant of positivism. It specifies that scientific concepts must be linked to instrumental procedures in order to determine their values. "Although a concept such as 'mass' may be conceived theoretically or metaphysically as a property, it is only pious opinion.... 'Mass' as a property is equivalent to 'mass' as inferred from pointer readings" (Blalock, 1961, p. 6).

In the social sciences, operationalism enjoyed a brief spell of acclaim through the work of sociologist George Lundberg (1939). He maintained that sociologists are wrong in believing that measurement can be

carried out only after things have been appropriately defined—for example, the idea that we must have a workable definition of alienation before we can measure it. For Lundberg, the definition comes about through measurement. He maintained that social values can be measured (and therefore defined) by an examination of the extent to which particular things are valued. Do people value X? If so, how much do they value X?

Operationalism remained fairly uncontroversial while the natural and social sciences were dominated by POSITIVISM but was an apparent casualty of the latter's fall from grace. In the social sciences, three main difficulties have been identified. First—and this is a problem shared with natural science—is that of underdetermination. How can the researcher know if testable propositions fully operationalize a theory? Concepts such as homelessness, poverty, and alienation take on a range of distinctive dimensions and appearances in different social environments. Homelessness, for instance, will depend for its definition on the historically and socially variable relationship between concepts of “shelter” and “home.” Second, the objective external definition of a phenomenon may not (or even cannot) accord with the subjective definition of those who experience it. What for one person seems to be poverty may be riches for another. The third problem is that of definitional disagreement between social scientists. Although individual researchers may find local practical solutions to the second problem, their solutions are likely to be different from and/or contradictory to those of other researchers.

These difficulties would appear to provide an undeniable refutation of the methodological approach, yet despite its apparent demise, operationalism remains an unresolved issue. All survey researchers are familiar with the process of *operationalization*, where one “descends the ladder of abstraction” from theory to measurement. This process is logically equivalent to that of operationalism. It is, for example, possible to state a broad, nonoperationalizable theory of homelessness, but if one wants to measure or describe homelessness using a social survey, then homelessness must be defined operationally (Williams, 1998). The difference between operationalization and operationalism perhaps lies in the relationship between theory and research. Operationalism leads to a division of labor between theory construction and measurement, but most survey researchers these days instead emphasize the importance of a dynamic relationship between

theory building and the specification of variables that can test a theory.

—Malcolm Williams

REFERENCES

- Blalock, H. (1961). *Causal inference in nonexperimental research*. Chapel Hill: University of North Carolina Press.
- Lundberg, G. (1939). *Foundations of sociology*. New York: Macmillan.
- Williams, M. (1998). The social world as knowable. In T. May & M. (Eds.), *Knowing the social world* (pp. 5–21). Buckingham, UK: Open University Press.

OPTIMAL MATCHING

Optimal matching (sometimes referred to as optimal alignment) is a metric technique for analyzing the overall resemblance between two strings or sequences of data. This approach makes it possible to compare sequences of uneven length or sequences composed of repeating elements. Combined with scaling or clustering ALGORITHMS, optimal matching can be used to detect empirically common sequential patterns across many cases.

The optimal matching approach was first developed by Vladimir Levenshtein in 1965 and since then has been widely used in the natural sciences for identifying similarity in DNA strings and in computer science as the basis of word- and speech-recognition algorithms. Andrew Abbott (1986) introduced the method to the social sciences.

Optimal matching rests on the premise that resemblance in sequential data can be quantified by calculating how easily one string can be turned into another. The transformation of one sequence into the other is accomplished via a set of elementary operators, usually including insertion of elements (I_i), deletion of elements (D_i), and substitution of one element for another (S_{ij}). Use of each of these operators is associated with a fixed penalty (which may vary for each element i or element pair ij). Using the costs associated with the operators, optimal matching algorithms identify the minimal total cost of transforming one sequence into another; this value is typically called the *distance* between the pair of sequences. Pairs of sequences with small distances resemble each other more than do pairs of sequences with larger distances.

S1	P	U	T	T	E	R	
S2	P	A	T	T	E	R	N

S1	P	∅	T	T	E	R	
		Â					Î
S2	P	A	T	T	E	R	N

Figure 1 Optimal Matching

If we consider words as sequences of letters, we can use optimal matching to calculate the resemblance between PUTTER and PATTERN (see Figure 1). If each operation exacts the same cost ($S_{ij} = I_i = D_i$ for all elements i and j), it is easy to see that in this example the simplest alignment involves two transformations: replacing the U in position 2 of PUTTER with an A, and inserting an N in position 8.

Although in some situations it may be reasonable to assign each transformation the same cost, most social sciences applications require a more complex cost scheme. For instance, optimal matching frequently is used to reveal patterns in career structures, where careers are built from jobs. Because some jobs are empirically more prevalent—and therefore more easily “substituted”—than are others, a common solution is to order the universe of elements (jobs) into a linear scale and then generate a MATRIX of substitution costs for each pair of elements by calculating the difference in their scores on the scale. Thus $S_{ij} = |i - j|$. In other substantive settings—such as sequences of collective action events or the history of policy changes—nonlinear substitution schemes with increasing or decreasing marginal costs may be more appropriate. Typically, insertion and deletion costs are equal and are set at a value close to the maximum substitution cost.

Although optimal matching is fundamentally a dyadic technique, it can be applied to large data sets with many cases of sequential data. In this situation, optimal matching algorithms produce a matrix in which each cell contains the distance between a pair of sequences. As with any distance matrix, scaling or clustering algorithms can be used to detect groups of cases that are close to one another; the interpretation

here is that small distances between cases reflect strong resemblance in the underlying sequential strings of data. Cases also can be classified in terms of their resemblance to theoretically derived or empirically typical sequences; such classifications can be used as VARIABLES in standard regression models.

—Katherine Stovel

REFERENCES

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, 16, 471–494.

Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data. *American Journal of Sociology*, 96, 144–185.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10, 707–710. (Original work published 1965.)

Sankoff, D., & Kruskal, J. B. (1983). *Time warps, string edits, and macromolecules*. Reading, MA: Addison-Wesley.

OPTIMAL SCALING

Optimal scaling, a term coined by R. Darrell Bock in 1960 for scaling of multivariate categorical data, is mathematically equivalent to DUAL SCALING, CORRESPONDENCE ANALYSIS, and homogeneity analysis of incidence data.

—Shizuhiko Nishisato

ORAL HISTORY

Oral history—research into the past that records the memories of witnesses to the past in order to draw on direct and personal experience of events and conditions—occupies an area of overlap between history, sociology, cultural studies, and psychology. Arguments in support of oral history turn on the challenge that it presents to official and dominant accounts of the past that are based in documentary sources. It is argued that oral history contributes to the broadening of the explanatory base and provides a “voice” to those whose experiences might otherwise go unrecorded. This emancipatory and critical thrust is, for many researchers, the attraction of the method (Perks & Thomson, 1998; Thompson, 2000). Critics question

the validity of DATA that depend on the reliability of lifetime survivors' memories.

THE INTERVIEW AND ITS ANALYSIS

At the heart of oral history is INTERVIEWING, a method of data collection that is central to social science practice but equally familiar to historians researching contemporary and 20th century history. For oral historians, the spoken immediacy of the interview, its social relations, and its inevitably interrogative nature are its defining characteristics, yet oral history remains a broad church. Depending on the disciplinary base of the oral historian and on the chosen topic, approaches to finding interviewees and interviewing may vary.

In an early and formative study, Paul Thompson (2000, pp. 145ff) used a QUOTA SAMPLE, taking a sociological approach to find subjects from populations that could not be subjected to randomized selection. When oral historians draw on late-life memories, they are working with survivors. In the Thompson case, a quota sample was used to fill predetermined categories with interview subjects, matching occupational groups identified in the 1911 U.K. census. Such an approach is favored when the focus is general social trends over time, for example, explorations of family relationships, gender, migration, employment, or political generations.

In such studies, interviews tend to be based on standardized areas for questioning, with topic areas identified, most frequently within a LIFE HISTORY perspective. Because the aim is to draw out comparable sources of evidence, to find regularities and patterns in what is presented, interviews tend to follow a closely prescribed form. Studies of this type are likely to be based on sample sizes nearer 100.

In contrast, studies that seek to explore particular events, or that draw on the experiences of a highly select group of survivors, or that focus on one individual's story tend to take a rather different approach to finding and interviewing subjects. In such cases, the most usual approaches to finding interviewees are SNOWBALL SAMPLING, appealing through the media, and using membership lists, representative bodies, or networks among older people's organizations to advertise for witness accounts.

Where studies draw on only a few accounts, the approach to interviewing may be rather different. For

one thing, there may be a return to interview the same person more than once, the aim being to secure a full account and one that will enable free expression of subjective experience on the part of the interviewee. Approaches vary again. For some oral historians, a more psychoanalytic, free-associative interaction has proved to be attractive. Explorations of memories of such events as childhood separation, experiences of segregation on the basis of race and disability, and strikes or conflicts can release powerful evocations, particularly when linked to other sources of evidence, documentary or visual, and when situated and explored in the interview as part of a whole life experience.

The identification of silences and suppression of experience has engaged the interest of oral historians whether or not they are influenced by psychoanalytic theories. Generally, among oral historians there has been a shift toward acceptance of more subjective, sometimes mythic accounts (Perks & Thomson, 1998). Recognizing that data such as these may have both cultural and psychological significance represents a move away from more empirically and positivistically situated approaches, with consequent implications for the interpretation of data and criticisms that focus on validity.

Analysis of interview data typically has followed the thematic approach used by both historians and sociologists with themes determined either in advance, according to the research design or the chosen theoretical framework, or deriving from the data itself with GROUNDED THEORY (Thompson, 2000, p. 151). Use of electronic methods has become more common, particularly with larger samples of interviews, though here again with an emphasis on the significance of language and expression. Researchers draw on the interpretive powers afforded by their particular disciplinary base—for example, history, psychology, sociology, or social policy—cross-referencing to other sources, both documentary and oral, seeking to locate and validate their analysis historically and structurally.

DEVELOPMENTS AND ISSUES

Oral history researchers, in common with developments in BIOGRAPHICAL METHODS and NARRATIVE ANALYSIS, anthropology, and ETHNOGRAPHY, have developed discussion relating to methods, significantly in the areas of FEMINIST RESEARCH, ethics, and ownership.

The position of women as subjects of research and as researchers has posed questions of power in the process of interviewing. The essentialist position argues that women share understandings and experience with their interviewees, giving particular resonance to their position as researchers. A contrasting position, equally feminist though more critical and reflective, acknowledges the salience of other aspects of difference in the interview relationship, notably class, color, and generation. A key contribution to these debates has been the edited volume by Sherna Berger Gluck and Daphne Patai (1991), which includes some searchingly self-critical accounts by feminists engaged in oral history research.

Ethical issues deriving from questions of consent and ownership now play a central role in oral history research. Under European Union law, copyright in the interview rests with the interviewee, which means that no part of the interview can be published unless copyright has been assigned to the interviewer or to his or her employing organization (Ward, 2002). Such a requirement is a legal expression of what some oral historians regard as central ethical issues, namely consent to participation in the process and ownership of what is recorded. Providing interviewees with control over the destination of the interview tape and transcript is regarded as a necessary stage in the interview process, with standard documentation now available on oral history Web sites and in key texts relating to ARCHIVE deposit and access.

—Joanna Bornat

REFERENCES

- Gluck, S. B., & Patai, D. (Eds.). (1991). *Women's words: The feminist practice of oral history*. London: Routledge.
- Perks, R., & Thomson, A. (Eds.). (1998). *The oral history reader*. London: Routledge.
- Thompson, P. (2000). *The voice of the past* (3rd ed.). Oxford, UK: Oxford University Press.
- Ward, A. (2002). *Oral history and copyright*. Retrieved June 7, 2002 from <http://www.oralhistory.org.uk>

ORDER

Order refers to the number of additional variables controlled or held constant in statistical analyses of a relationship between two variables. A zero-order relationship measures the magnitude or strength

of an association between two variables, without controlling for any other factors (Knoke, Bohrnstedt, & Mee, 2002, p. 213). Familiar examples include CONTINGENCY TABLES for two categorical variables, CORRELATION between two continuous variables, and simple REGRESSION of a dependent variable on one independent variable.

A first-order relationship involves controlling for the effects of a third variable to determine whether a bivariate relationship changes. In cross-tabulation analysis, this procedure is called ELABORATION (Kendall & Lazarsfeld, 1950). A relationship between variables X and Y is elaborated by splitting their zero-order cross-tabulation into first-order subtables, also called partial tables. Each subtable exhibits the $X - Y$ association within one category of control variable Z . If the $X - Y$ association is not affected by Z , then measures of association such as PHI COEFFICIENT or GAMMA will have identical values in each first-order table. If the association is spurious because Z causes both variables, then the association disappears in each subtable; hence, Z "explains" the $X - Y$ relationship. The same outcome occurs if Z intervenes in a causal sequence linking X to Y ; hence, Z "interprets" the $X - Y$ relationship. Elaboration often uncovers a "specification" or INTERACTION EFFECT, where the strength or direction of association differs across subtables, indicating that the $X - Y$ relationship is contingent or conditional on the specific categoric of Z .

A first-order relationship for three continuous variables can be estimated using the PARTIAL CORRELATION coefficient, which shows the "magnitude and direction of covariation between two variables that remain after the effects of a control variable have been held constant" (Knoke et al., 2002, p. 223). Its formula is a function of three zero-order correlations:

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{1 - r_{XZ}^2} \sqrt{1 - r_{YZ}^2}}.$$

If X and Y are uncorrelated with variable Z , their partial correlation equals their zero-order correlation, $r_{XY \cdot Z} = r_{XY}$; that is, their linear relationship is unaffected by the control variable. However, if Z correlates strongly in the same direction with both variables, their reduced (or zero) partial correlation reveals that some of or all the $X - Y$ relationship is spurious.

Second- and higher-order relationships generalize these methods for analyzing zero-order relations using multiple control variables. LOG-LINEAR MODELING

estimates complex interaction effects in multivariate cross-tabulations. In MULTIPLE REGRESSION ANALYSIS equations, each partial regression coefficient shows the effect on a dependent variable of one-unit differences in an independent variable after controlling for ("partialing out") additive effects of other predictor variables. Detecting INTERACTION EFFECTS involves forming product terms or other independent variable combinations (Jaccard, Turrisi, & Wan, 1990).

Order also refers to a MOMENT of the distribution of a random variable X . Moments about the origin are defined for the k th positive integer as the expected value of X^k , or $E(X^k)$; thus, the first moment about the origin, $E(X)$, is the MEAN. If the mean is subtracted from X before exponentiation, the moments about the mean are defined as $E[X - E(X)]^k$. The first moment of X about its mean equals zero, while its second moment is the VARIANCE. The third and fourth moments about the mean are called, respectively, skewness and KURTOSIS.

—David H. Knoke

REFERENCES

- Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. Newbury Park, CA: Sage.
- Kendall, P. L., & Lazarsfeld, P. F. (1950). Problems of survey analysis. In R. K. Merton & P. Lazarsfeld (Eds.), *Continuities in social research* (pp. 133–196). New York: Free Press.
- Knoke, D., Bohrnstedt, G. W., & Mee, A. P. (2002). *Statistics for social data analysis* (4th ed.). Itasca, IL: Peacock.

ORDER EFFECTS

Question and response-order effects in SURVEY interviews and QUESTIONNAIRES have become the focus of numerous methodological experiments by sociologists, political scientists, public opinion researchers, and cognitive social psychologists. Just as the meaning of a word or phrase in spoken and written communication cannot be separated from the context in which it occurs, so too have researchers discovered that the meaning of a survey question can be influenced significantly by the respondent's reactions to previous questions in an INTERVIEW or SELF-ADMINISTERED QUESTIONNAIRE. They also have learned that order effects can occur within, as well as between, survey questions because of the serial presentation of response alternatives. Respondents evidently use the order of the questions and the response categories to make

inferences about the meaning of the questions asked of them.

Question order and context effects often arise when respondents are asked two or more questions about the same subject or about closely related topics. They can show up even if multiple items in the interview or questionnaire separate closely related questions. Initial attempts to understand such effects have included the development of conceptual classifications of the types of relations among questions: (a) part-whole relationships involving two or more questions, in which one of them is more general than the other and therefore implies or subsumes the other question, but not necessarily the reverse (e.g., how happy one is with life in general vs. one's marital happiness); and (b) part-part relationships involving questions asked at the same level of specificity (e.g., permitting an abortion in the case of rape vs. a case of a potential birth defect). Two types of effects have likewise been identified: (a) consistency effects, in which respondents bring answers to subsequent questions more in line with their answers to earlier questions (e.g., voting preferences and their political party identification); and (b) contrast effects, in which they differentiate their responses to later questions from those given to previous questions (e.g., their approval of the way George W. Bush has handled the economy vs. how he has handled foreign relations). The net result is a fourfold typology of question order effects: part-part consistency effects, part-whole contrast effects, part-part contrast effects, and part-whole consistency effects (Schuman & Presser, 1981, chap. 2). Smith (1992) later demonstrated that order effects depend not only on the holistic context of previous questions, as identified in Schuman and Presser's typology, but also on how respondents actually answered the prior questions. Conceptually, this interaction between question order and responses to antecedent questions has become known as "conditional order effects." More recent psychological research has used the heuristic of cognitive accessibility, models of belief sampling, the logic of conversation, and cognitive models of assimilation and contrast effects to make theoretical sense of various types of question order and context effects (see, e.g., Sudman, Bradburn, and Schwarz, 1996).

Response-order effects within survey questions can appear in several ways. When the response categories for a question are read aloud, as in a telephone interview or in a face-to-face interview without a visible response card, respondents will tend to select

the last-mentioned response alternative: a *recency effect*. If the response categories are presented in a visual format, as in a self-administered mail survey, Web-based poll, or exit poll, respondents will tend to choose one of the first listed alternatives: a *primacy effect*. Judgmental *contrast effects* among response categories can arise as well when a questionnaire item is preceded by another item that generates more extreme ratings on some response scale dimension, as for example when an extremely well-liked or disliked political figure is presented in the middle of a list of other figures to be rated. Explanations of response-order effects have ranged from the conventional opinion crystallization hypothesis, which predicts that such effects will be less likely to occur among respondents with intensely held prior opinions on the topic; to the memory limitations hypothesis, which postulates that such effects occur because of the inability of respondents to remember all the response alternatives; to more current cognitive elaboration and satisficing models (see, e.g., the review by Bishop and Smith, 2001). Derived from the literature on attitude persuasion in cognitive social psychology, the cognitive elaboration model assumes that the mediating process underlying response-order effects is the opportunity respondents have to think about (elaborate on) the response alternatives presented to them. When a list of response alternatives is read aloud to respondents, as for example in a TELEPHONE SURVEY, respondents have more opportunity to think about (elaborate on) choices presented to them later in the list, which often generates a recency effect. In contrast, the satisficing model rooted in Herbert Simon's theory of economic decision making, assumes that most survey respondents answer survey questions by selecting the first satisfactory or acceptable response alternative offered to them rather than taking the time to select an optimal answer. Selecting the last-mentioned category in a telephone survey question, for example, demonstrates the tendency of respondents to satisfice rather than optimize. Despite these promising cognitive theoretical developments, both question and response-order effects have continued to resist ready explanation and prediction.

—George F. Bishop

REFERENCES

- Bishop, G., & Smith, A. (2001). Response-order effects and the early Gallup split-ballots. *Public Opinion Quarterly*, 65, 479–505.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.
- Smith, T. W. (1992). Thoughts on the nature of context effects. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 163–184). New York: Springer-Verlag.
- Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

ORDINAL INTERACTION

Distinctions often are made between *ordinal* interactions and *disordinal* interactions. The distinction usually is made in the context of REGRESSION analyses, where one is comparing the slope of Y on X for one group with the slope of Y on X for another group. An INTERACTION is implied if the slopes for the two groups are not equal. For each group, one can specify a regression line that characterizes the relationship between Y and X across the range of X values observed in the groups. Nonparallel regression lines for the two groups imply the presence of interaction. A disordinal interaction is one in which the regression line that regresses Y onto X for one group intersects the corresponding regression line for the other group. This also is referred to as a *crossover* interaction. An ordinal interaction is one in which the regression lines are nonparallel, but they do not intersect within the range of data being studied.

For any given pair of nonparallel regression lines, there is always a point where the lines will intersect. In this sense, all interactions are disordinal in theory. Interactions are classified as being ordinal if, *within the range of scores being studied* (e.g., for IQ scores between 90 and 110), the regression lines do not intersect.

If one knows the values of the INTERCEPT and SLOPE for the two regression lines, the value of X where the two regression lines intersect can be calculated algebraically. Let $\beta_{0,A}$ be the intercept for group A, $\beta_{1,A}$ be the slope for group A, $\beta_{0,B}$ be the intercept for group B, and $\beta_{1,B}$ be the slope for group B. The point of intersection is

$$PI = (\beta_{0,A} - \beta_{0,B}) / (\beta_{1,B} - \beta_{1,A}).$$

Applied researchers express some wariness about ordinal interactions. Such interactions, they contend, may be an artifact of the nature of the metric of the outcome variable. Nonparallel regression lines

frequently can be made parallel by simple MONOTONIC transformations of scores on the outcome variable. If the metric of the outcome variable is arbitrary, then some scientists argue that one should consider removing the interaction through transformation in the interest of MODEL parsimony. Ordinal interactions, however, should not be dismissed if the underlying metric of the variables is meaningful.

—James J. Jaccard

ORDINAL MEASURE

A measure is a data point. In many research applications, a measure is a number assigned to an individual, organism, or research participant that reflects his or her or its standing on some construct. A measure is taken at a given time, on a given individual, in a given setting, using a given type of recording device. A set of measures refers to multiple data points. For example, a measure of height on each of 10 individuals represents a set of 10 measures.

Ordinal measures can be described from different vantage points. Ordinality as applied to a set of measures describes a property of the function that relates observed measures of a CONSTRUCT to the true values of that construct. A set of measures of a construct has the property of ordinality if the values of the measure are a strictly MONOTONIC function of the true values of the construct, that is, if the relative ordering of individuals on the true construct is preserved when one uses the observed measure to order individuals. For example, 10 individuals may differ in their height, and they can be ordered from shortest to tallest. A researcher may develop a strategy for measuring height such that higher scores on the index imply greater height. The set of measures taken on the 10 individuals is said to have ordinal properties if the ordering of individuals on the index is the same as the ordering of individuals on the underlying dimension of height.

Strictly speaking, ordinal measures do not convey information about the magnitude of differences between individuals or organisms on the underlying dimension. They only permit us to state that one individual has more or less of the dimension than another individual.

—James J. Jaccard

REFERENCE

Anderson, N. H. (1981). *Methods of information integration theory*. New York: Academic Press.

ORDINARY LEAST SQUARES (OLS)

Consider a linear relationship in which there is a stochastic dependent variable with values y_i that are a function of the values of one or more nonstochastic independent variables, $x_{1i}, x_{2i} \dots x_{ki}$. We could write such a relationship as $y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i$. For any such relationship, there is a *prediction equation* that produces predicted values of the dependent variable. For example, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_k x_{ki}$. Predicted values usually are designated by the symbol $\hat{y}_i$. If we subtract the predicted values from the actual values ($y_i - \hat{y}_i$), then we produce a set of RESIDUALS (ε_i), each of which measures the distance between a prediction and the corresponding actual observation on y_i . Ordinary Least Squares (OLS) is a method of point estimation of parameters that minimizes the function defined by the sum of squares of these residuals (or distances) with respect to the parameters.

OLS ESTIMATION

For example, consider the simple bivariate regression given by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The prediction equation is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The residuals can be calculated as $\varepsilon_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = y_i - \hat{y}_i$. Then the OLS estimator is defined by finding the minimum of the function $S = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ with respect to the parameters β_0 and β_1 . This is a concave quadratic function that always has a single minimum. This simple minimization problem can be solved analytically. The standard approach to minimizing such a function is to take first partial derivatives with respect to the two parameters β_0 and β_1 , then solve the resulting set of two simultaneous equations in two unknowns. Performing these operations, we have the following:

$$\begin{aligned} & \frac{\partial \left[\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right]}{\partial \beta_0} \\ &= \sum_{i=1}^n 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0 \end{aligned}$$

$$\frac{\partial \left[\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right]}{\partial \hat{\beta}_1} = \sum_{i=1}^n 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0.$$

Simplifying this result and rearranging produces what are called the normal equations:

$$\sum_{i=1}^n y_i = n\hat{\beta}_0 + \left(\sum_{i=1}^n x_i \right) \hat{\beta}_1,$$

$$\sum_{i=1}^n x_i y_i = \left(\sum_{i=1}^n x_i \right) \hat{\beta}_0 + \left(\sum_{i=1}^n x_i^2 \right) \hat{\beta}_1.$$

Because we can easily calculate the sums in the preceding equations, we have only to solve the resulting set of two simultaneous equations in two unknowns to have OLS estimates of the regression parameters. These analytical procedures are easily extended to the multivariate case, resulting in a set of normal equations specific to the dimension of the problem. For example, see Greene (2003, p. 21).

CLASSICAL ASSUMPTIONS FOR OLS ESTIMATION

The OLS estimator is incorporated into virtually every statistical software package and is extremely popular. The reason for this popularity is its ease of use and very convenient statistical properties. Recall that *parameter estimation* is concerned with finding the value of a population parameter from sample statistics. We can never know with certainty that a sample statistic is representative of a population parameter. However, using well-understood statistical theorems, we can know the properties of the estimators used to produce the sample statistics.

According to the GAUSS-MARKOV THEOREM, an OLS regression estimator is a BEST LINEAR UNBIASED ESTIMATOR (BLUE) under the following assumptions.

1. The expected value of the population disturbances is zero ($E[\varepsilon_i] = 0$). This assumption implies that any variables omitted from the population regression function do not result in a systematic effect on the mean of the disturbances.
2. The expected value of the covariance between the population independent variables and disturbances is zero ($E[X_i, \varepsilon_i] = 0$).

This assumption implies that the independent variables are nonstochastic, or, if they are stochastic, they do not covary systematically with the disturbances.

3. The expected value of the population correlation between disturbances is zero ($E[\varepsilon_i, \varepsilon_j] = 0$). This is the so-called nonauto-correlation assumption.
4. The expected value of the population variance of the disturbances is constant, ($E[\varepsilon_i^2] = 0$). This is the so-called nonheteroskedasticity assumption.

These four assumptions about the population disturbances are commonly called the Gauss-Markov assumptions. Other conditions, however, are relevant to the statistical properties of OLS. If disturbances are normal, then the OLS estimator is the best unbiased estimator (BUE), but if disturbances are non-normal, then the OLS estimator is only the best linear unbiased estimator (BLUE). That is, another estimator that is either nonlinear or robust may be more efficient depending on the nature of the non-normality.

Additionally, the population regression function must be correctly specified by the sample regression function. All relevant independent variables should be included, and no nonrelevant independent variables should be included. The functional form of the population regression function should be linear.

Another requirement of the OLS estimator is the absence of perfect MULTICOLLINEARITY. Perfect multicollinearity exists when one of the independent variables is a perfect linear combination of one or more other independent variables. If perfect multicollinearity exists, then the OLS estimator is undefined and the standard error of the sampling distribution is infinite. High multicollinearity may be problematic for enabling the researcher to disentangle the independent effects of the regressors. However, even in the presence of high multicollinearity, the OLS estimator is BLUE.

STATISTICAL INFERENCE USING OLS ESTIMATORS

Knowing that the OLS estimator is BLUE provides absolute certainty that *on average*, under repeated sampling, the estimated sample statistic will equal the population parameter. Additionally, the variance of sample statistics under repeated sampling when using OLS will be the smallest possible among all possible estimators of this variance.

Although we can never be certain that the sample statistic for any particular sample is representative of the population, knowledge of the estimator that produced the sample statistic enables assessing the degree of confidence we can have about estimates. For example, the standard deviation of the OLS sampling distribution of regression slope coefficients is given as

$$SE(\hat{\beta}_k) = \sqrt{\frac{\hat{\sigma}_e^2}{\sum_{i=1}^n (X_i - \bar{X})^2 (1 - R_j^2)}},$$

where R_j^2 is the explained variance from the regression of variable j on all other independent variables. This is also referred to as the standard error. Using the standard error, we can make probability statements about the likelihood that a sample statistic is representative of a population parameter. Relying on knowledge that the distribution of the sample statistic under repeated sampling is normal, the probability of an estimate being more than s standard deviations away from the true parameter can be calculated easily. More often, however, the t -distribution is used to take into account the fact that the disturbance variance in the numerator of the preceding equation is an estimate, rather than the true population variance.

Although not technically a requirement for OLS estimation, the assumption of normally distributed population disturbances usually is made. The consequence of non-normal disturbances, however, is quite serious. Non-normality of the “fat-tailed” kind means that hypothesis tests cannot be done meaningfully. In this situation, there are two options. First, the data often can be transformed to produce normally distributed disturbances. Second, one can employ one of the available robust estimators that takes into account the non-normally distributed disturbances.

—B. Dan Wood and Sung Ho Park

REFERENCES

- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.
- Judge, G. G., Hill, R. C., Griffiths, W. E., Lutkepohl, H., & Lee, T. (1988). *Introduction to the theory and practice of econometrics* (2nd ed.). New York: John Wiley & Sons.
- Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). *Understanding regression analysis: An introductory guide*. Beverly Hills, CA: Sage.

ORGANIZATIONAL ETHNOGRAPHY

Dating from the 19th century, ETHNOGRAPHY (“nation-description”) has meant the portrayal of a people: their customs, manners, and beliefs. Organizational ethnography emerged in the 1980s with the growing interest in culture, symbolism, and folklore. Through the use of FIELD RESEARCH, ethnographers document traditions and symbolic behaviors that reveal people’s feelings, opinions, and customary ways of doing things as well as the ethos of an organization. Goals include understanding organizational behavior and contributing to organization development.

The early 1980s witnessed a search for new paradigms with which to think about organizations: perspectives that would direct attention toward the human dimension in ways that mechanistic and super-organic models did not. In 1980, the *Academy of Management Review* published an article by Dandridge, Mitroff, and Joyce on “organizational symbolism.” A special issue of *Administrative Science Quarterly* appeared in 1983 devoted to the subject of “organizational culture.” A conference directed by Michael Owen Jones that same year titled “Myth, Symbols & Folklore: Expanding the Analysis of Organizations” introduced the notion of “organizational folklore.” Although John Van Maanen (1979) employed the term *organizational ethnography* earlier, probably Evergreen State College in Olympia, Washington, should receive credit as the first institution to publicly seek one. In May, 1990, it advertised for an ethnographer to uncover its culture, selecting Peter Tommerup (1993), a PhD candidate in folklore and mythology. Through interviews and observations focusing on stories, rituals, and other expressive forms (Georges & Jones, 1995), he gathered data to describe the nature and assess the impact of the teaching and learning cultures at the college.

Organizational ethnography relies on in-depth interviewing, observation, and PARTICIPANT OBSERVATION in order to capture the “feel” of the organization from the inside, the perceptions of individual members, and the community of shared symbols (Jones, 1996). Researchers elicit participants’ accounts of events in their own words and expressive media, whether story, celebration, ritual, proverb, or objects. They note the personal decoration of workspace, graffiti, poster and flyer art, photographs, bulletin boards, images and

sayings posted on people's doors, photocopy lore, and other material manifestations of behavior and culture. They record verbal expressions such as traditional sayings, metaphors, slogans, oratory, rumors, beliefs, personal experience stories, and oft-told tales. They observe (and sometimes participate in) customs, rituals, rites of passage, festive events, play, gestures, social routines, and ceremonies.

By documenting and analyzing examples of organizational symbolism, ethnographers gain insights into leadership style, climate, ambiguity and contradictions, tensions and stress, conflicts, and coping in organizations. Ethnography also brings to light values and ways of doing things long taken for granted that guide people's actions, inform decision making, and aid or hinder organizational effectiveness.

—Michael Owen Jones

REFERENCES

- Dandridge, T. C., Mitroff, I., & Joyce, W. F. (1980). Organizational symbolism: A topic to expand organizational analysis. *Academy of Management Review*, 5, 77–82.
- Georges, R. A., & Jones, M. O. (1995). *Folkloristics: An introduction*. Bloomington: Indiana University Press.
- Jones, M. O. (1996). *Studying organizational symbolism: What, how, why?* Thousand Oaks, CA: Sage.
- Tommerup, P. (1993). *Adhocratic traditions, experience narratives and personal transformation: An ethnographic study of the organizational culture and folklore of the Evergreen State College, an innovative liberal arts college*. Unpublished doctoral dissertation, University of California, Los Angeles.
- Van Maanen, J. (1979). The fact and fiction in organizational ethnography. *Administrative Science Quarterly*, 24, 539–550.

ORTHOGONAL ROTATION.

See ROTATIONS

OTHER, THE

The *other* is a term used in human science discourses that are concerned with themes such as identity, difference, selfhood, recognition, and ethics. *Other* can mean the other person but also the self as other. Philosophy has a long-standing interest in addressing

the notion of the other but often has done so by reducing the “other” to the “same” or the “self.” Self-discovery or self-recognition can be seen as a process of becoming, whereby the self recognizes itself in the alterity (otherness) of objects. For example, Hegel showed how self-consciousness and freedom can be achieved only through the overcoming of what is other.

Much of contemporary human science literature dealing with the notion of other finds its impetus in the early phenomenological writings of Emmanuel Levinas. Here, the other transcends the self and cannot be reduced to the self. For Levinas, the otherness of the other is first of all an ethical event and appears to us in the nakedness and vulnerability of the face. When we have a Levinassian encounter with a person who is weak or vulnerable, then this other makes an appeal on us that we have already experienced as an ethical response before we are reflectively aware of it. “From the start, the encounter with *other* is my responsibility for him,” said Levinas (1998, p. 103).

In her classic book *The Second Sex* (1949/1970), Simone de Beauvoir examined by means of historical, literary, and mythical sources how the oppression of women is associated with patriarchal processes of systematic objectification. The result is that the male is seen as “normal” and the female as “other,” leading to the loss of women's social and personal identity. In Kate Millett's *Sexual Politics* (first published in 1970, with revisions afterward) and in the work of contemporary feminists, women consistently appear as “other” against male constructions of social norms.

French poststructuralists in particular have developed a range of interpretations of the notions of other, otherness, and othering. Othering is what people do when they consider themselves normal but notice differences in other peoples. In the works of Michel Foucault (1995), others are those who have been victimized and marginalized—the mentally ill, prisoners, women, gays, lesbians, and people of color among them. Genealogical analysis consists of recovering the social, cultural, and historical roots of the practices of institutionalization and marginalization of the ones who are our others.

Paul Ricoeur has explored the dialectical ties between selfhood and otherness in his study of personal and narrative identity. He is especially interested in the experience of otherness that resides in the body of the self. There are several levels of otherness, said Ricoeur. There is the otherness of “myself as flesh,” the otherness of the other person (the foreigner), and the otherness experienced in “the call” of conscience. These

distinctions may help us to understand the programs of inquiry of early and contemporary scholars.

In Julia Kristeva's *Strangers to Ourselves* (1991), the notion of the other refers to the foreigners, outsiders, or aliens around us. She asks how we can live in peace with these foreigners if we cannot come to terms with the other inside ourselves. In her analysis she aims to show how alterity of the other is the hidden face of our own identity. Poststructuralist studies of the other in intersubjective relations is further exemplified in the influential work of Jacques Derrida where the other is a theme for raising the question of responsibility, duty, and moral choice. He said, "I cannot respond to the call, the request, the obligation, or even the love of another without sacrificing the other other, the other others" (Derrida, 1995, p. 68).

The question of the meaning of other and otherness figures prominently in qualitative inquiries in fields such as multiculturalism, gender studies, queer theory, ethnicity, and postcolonial studies. In writings on the politics of multiculturalism by Charles Taylor, the "significant other" is addressed in terms of the searches for self-identity and complex needs for social recognition that are sometimes incommensurable. In philosophy too, the notion of other has become a common concern. Especially worth mentioning are the phenomenological studies of the other by Alphonso Lingis, who translated some of the main works of Levinas. In his book *The Community of Those Who Have Nothing in Common* (1994b), Lingis started an inquiry into the question how we can experience moral responsibility and community with respect to cultural others with whom we have no racial kinship, no language, no religion, and no economic interests in common. He continued this line of research in studies such as *Abuses* (1994a) and *The Imperative* (1998) that constitute an entirely new approach to philosophy. In his work, Lingis has blended evocative methods of travelogue, letter writing, ethics, and meditations on themes such as torture, war, identity, lust, passions, and the rational in his "face to face" encounters with marginalized others in India, Bali, Bangladesh, and Latin America.

—Max van Manen

REFERENCES

- de Beauvoir, S. (1970). *The second sex*. New York: Alfred A. Knopf. (Original work published 1949.)
- Derrida, J. (1995). *The gift of death*. Chicago: The University of Chicago Press.
- Foucault, M. (1995). *Discipline and punish: The birth of the prison*. New York: Vintage Books.
- Kristeva, J. (1991). *Strangers to ourselves*. New York: Columbia University Press.
- Levinas, E. (1998). *On thinking-of-the-Other: Entre nous*. New York: Columbia University Press.
- Lingis, A. (1994a). *Abuses*. Berkeley: The University of California Press.
- Lingis, A. (1994b). *The community of those who have nothing in common*. Bloomington: Indiana University Press.
- Lingis, A. (1998). *The imperative*. Bloomington: Indiana University Press.
- Millett, K. (1970). *Sexual politics*. Garden City, NY: Doubleday.
- Ricoeur, P. (1992). *Oneself as another*. Chicago: The University of Chicago Press.
- Taylor, C. (1994). The politics of recognition. In C. Taylor & A. Gutmann (Eds.), *Multiculturalism and the politics of recognition* (pp. 25–73). Princeton, NJ: Princeton University Press.

OUTLIER

An outlier is an OBSERVATION whose value falls very far from the other values. For example, in a community income study, suppose there is a millionaire with an income of \$2,000,000, well beyond the next highest income of \$150,000. That millionaire is a outlier on the income variable. Outliers can pose problems of inference, especially if they are extremely far out, or if there are very many of them. For example, in an income study of community leaders, $N = 100$, where a handful of them are multimillionaires, the estimated mean income would give a distorted view of the income of the typical community leader. A better measure of central tendency, in this instance, would be the MEDIAN. Outliers also can distort other statistics, such as the REGRESSION COEFFICIENT. A common remedy is to exclude them from analysis, but that approach is flawed in that it loses information. TRANSFORMATION of the outliers may be the ideal solution, for it can pull these extreme values in, at the same time maintaining the full sample.

—Michael S. Lewis-Beck

See also INFLUENTIAL STATISTICS

P

***P* VALUE**

The p value for a statistical test of a hypothesis is defined as the PROBABILITY, calculated under the assumption that the NULL HYPOTHESIS is true, of obtaining a sample result as extreme as, or more extreme than, that observed in a particular direction. This direction is determined by the direction of the ONE-TAILED (or ONE-SIDED) alternative. As an example, consider the SIGN TEST for the null hypothesis that a coin is balanced. The coin is tossed 10 times, and we observe heads 9 times. The sample size is $n = 10$, and the test statistic is S , defined as the number of heads in this sample. The observed sample result is $S = 9$. This sample result is a fact. The question is whether this sample result is consistent with the claim (in the null hypothesis) that the coin is balanced. If it is not, then the claim must be in error, and hence the null hypothesis should be rejected. To determine whether the sample result is consistent with this claim, we calculate the probability of obtaining 9 or more heads in a sample of 10 tosses when the coin is balanced. This probability is found from the binomial distribution with $n = 10$ and $p = 0.50$ for the probability of a head when the coin is balanced, or

$$\begin{aligned} p \text{ value} &= p(S \geq 9) = \sum_{t=9}^{10} \binom{10}{t} (0.50)^t (0.50)^{10-t} \\ &= 0.0020. \end{aligned}$$

This represents a 2 in 1,000 chance, which is quite a rare occurrence. Therefore, our decision is to

reject the null hypothesis and conclude that the coin is not balanced. The coin seems to be weighted in favor of the occurrence of a head over a tail; that is, $p > 0.50$.

The traditional or classical approach to this hypothesis-testing problem is to preselect a numerical value for α , the probability of a TYPE I ERROR, and use that α to determine a critical region or rejection region for the value of S ; for example, reject if $S \geq c$ for some constant c . This preselected α value is almost always completely arbitrary (0.05 may be traditional, but it is totally arbitrary), but it can have a tremendous effect on the decision that is reached.

Reporting the p value is an alternative method of carrying out a test of a null hypothesis against a one-sided alternative. Instead of reaching a definitive decision to reject or not to reject the null hypothesis at a chosen level α , the researcher lets the readers or users make their own decision based on the magnitude of this p value. The readers can choose their own α . Then the decision made by the readers is to reject if $p \leq \alpha$ and not to reject if $p > \alpha$.

If the ALTERNATIVE HYPOTHESIS does not predict a particular direction—that is, if the alternative is TWO-TAILED (or TWO-SIDED)—then the p value is not uniquely defined. There are essentially two different p values, one in each possible direction. The smaller of these two possible p values is the only relevant one. Then the researcher can either report this smaller p value, indicating that it is a one-sided p value, or double this one-sided p value.

—Jean D. Gibbons

PAIRWISE CORRELATION. See CORRELATION

PAIRWISE DELETION

Pairwise deletion is a term used in relation to computer software programs such as SPSS in connection with the handling of MISSING DATA. Pairwise deletion of missing data means that only cases relating to each pair of variables with missing data involved in an analysis are deleted. Consider the following scenario: We have a sample of 200 cases and want to produce a set of PEARSON CORRELATIONS for 10 variables. Let us take the first 3 variables. Variable 1 has 4 missing cases, Variable 2 has 8 missing cases, and Variable 3 has 2 missing cases. Let us also assume that the missing cases are different for each of the variables; that is, Variable 1's 4 missing cases are not among Variable 2's 8 missing cases. (In practice, this will not always be the case because if a person does not answer questions relating to both Variable 1 and Variable 2, he or she will be a case with missing data on both variables.) When we analyze the 10 variables, the correlation between Variables 1 and 2 will be based on 188 cases (because between them, Variable 1 and Variable 2 have 12 missing cases). The correlation between Variable 1 and Variable 3 will be based on 194 cases (because between them, they have 6 missing cases), and between Variable 2 and Variable 3, it will be based on 190 cases (because between them, they have 10 missing cases).

—Alan Bryman

See also DELETION, LISTWISE DELETION

PANEL

The term *panel* refers to a RESEARCH DESIGN in which the DATA are gathered on at least two occasions on the same units of analysis. Most commonly, that unit of analysis is an individual in a SURVEY. In a two-wave panel survey, respondents at time t are reinterviewed at time $t + 1$; in a three-wave panel, they would be reinterviewed yet again at time $t + 2$.

Sometimes, the units of analysis of the panel are aggregates, such as nations. For example, a sample of European countries might be measured at time t and again at $t + 1$. A special value of a panel design is that it incorporates the temporality required for strong CAUSAL inference because the potential causal variables, the X s, can actually be measured before Y occurs. In addition, panel studies allow social change to be gauged. For example, with repeated INTERVIEWING over a long time, panel surveys (unlike COHORT designs) have the potential to distinguish age effects, PERIOD EFFECTS, and cohort effects.

A chief disadvantage of the panel approach is the data ATTRITION from time point to time point, especially with panel surveys in which individuals drop out of the sample. A related problem is the issue of addition to the panel, in that demographic change may, over time, suggest that the original panel is no longer representative of the POPULATION initially SAMPLED. For example, as a result of an influx of immigrants, it may be that the panel should have respondents added to it to better reflect the changed population.

—Michael S. Lewis-Beck

PANEL DATA ANALYSIS

Panel data refer to data sets consisting of multiple observations on each sampling unit. This could be generated by pooling time-series observations across a variety of cross-sectional units, including countries, states, regions, firms, or randomly sampled individuals or households. This encompasses longitudinal data analysis in which the primary focus is on individual histories. Two well-known examples of U.S. panel data are the Panel Study of Income Dynamics (PSID), collected by the Institute for Social Research at the University of Michigan, and the National Longitudinal Surveys of Labor Market Experience (NLS) from the Center for Human Resource Research at Ohio State University. An inventory of national studies using panel data is given at <http://www.ceps.lu/Cher/Cherpres.htm>. These include the Belgian Household Panels, the German Socio-economic Panel, the French Household Panel, the British Household Panel Survey, the Dutch Socio-economic Panel, the Luxembourg Household Panel, and, more recently, the European Community household panel. The PSID began in 1968 with 4,802

families and includes an oversampling of poor households. Annual interviews were conducted and socioeconomic characteristics of each family and roughly 31,000 individuals who had been in these or derivative families were recorded. The list of variables collected is more than 5,000. The NLS followed five distinct segments of the labor force. The original samples include 5,020 men ages 45 to 59 years in 1966, 5,225 men ages 14 to 24 years in 1966, 5,083 women ages 30 to 44 years in 1967, 5,159 women ages 14 to 24 years in 1968, and 12,686 youths ages 14 to 21 years in 1979. There was an oversampling of Blacks, Hispanics, poor Whites, and military in the youths survey. The variables collected run into the thousands. Panel data sets have also been constructed from the U.S. Current Population Survey (CPS), which is a monthly national household survey conducted by the Census Bureau. The CPS generates the unemployment rate and other labor force statistics. Compared with the NLS and PSID data sets, the CPS contains fewer variables, spans a shorter period, and does not follow movers. However, it covers a much larger sample and is representative of all demographic groups.

Some of the benefits and limitations of using panel data are given in Hsiao (1986). Obvious benefits include a much larger data set because panel data are multiple observations on the same individual. This means that there will be more variability and less COLLINEARITY among the variables than is typical of cross-section or time-series data. For example, in a demand equation for a given good (say, gasoline) price and income may be highly correlated for annual time-series observations for a given country or state. By stacking or pooling these observations across different countries or states, the variation in the data is increased and collinearity is reduced. With additional, more informative data, one can get more reliable estimates and test more sophisticated behavioral models with less restrictive assumptions. Another advantage of panel data is their ability to control for individual heterogeneity. Not controlling for these unobserved individual specific effects leads to BIAS in the resulting estimates. For example, in an earnings equation, the wage of an individual is regressed on various individual attributes, such as education, experience, gender, race, and so on. But the error term may still include unobserved individual characteristics, such as ability, which is correlated with some of the regressors, such as education. Cross-sectional studies attempt to control for this unobserved ability by collecting hard-to-get

data on twins. However, using individual panel data, one can, for example, difference the data over time and wipe out the unobserved individual invariant ability. Panel data sets are also better able to identify and estimate effects that are not detectable in pure cross-section or pure time-series data. In particular, panel data sets are better able to study complex issues of dynamic behavior. For example, with cross-section data, one can estimate the rate of unemployment at a particular point in time. Repeated cross-sections can show how this proportion changes over time. Only panel data sets can estimate what proportion of those who are unemployed in one period remains unemployed in another period.

Limitations of panel data sets include the following: problems in the design, data collection, and data management of panel surveys (see Kasprzyk, Duncan, Kalton, & Singh, 1989). These include the problems of coverage (incomplete account of the population of interest), nonresponse (due to lack of cooperation of the respondent or because of interviewer error), recall (respondent not remembering correctly), frequency of interviewing, interview spacing, reference period, the use of bounding to prevent the shifting of events from outside the recall period into the recall period, and time-in-sample bias. Another limitation of panel data sets is the distortion due to measurement errors. Measurement errors may arise because of faulty response due to unclear questions, memory errors, deliberate distortion of responses (e.g., prestige bias), inappropriate informants, misrecording of responses, and interviewer effects. Although these problems can occur in cross-section studies, they are aggravated in panel data studies. Panel data sets may also exhibit bias due to sample selection problems. For the initial wave of the panel, respondents may refuse to participate, or the interviewer may not find anybody at home. This may cause some bias in the inference drawn from this sample. Although this nonresponse can also occur in cross-section data sets, it is more serious with panels because subsequent waves of the panel are still subject to nonresponse. Respondents may die, move, or find that the cost of responding is high. The rate of attrition differs across panels and usually increases from one wave to the next, but the rate of increase declines over time. Typical panels involve annual data covering a short span of time for each individual. This means that asymptotic arguments rely crucially on the number of individuals in the panel tending to infinity. Increasing the time span of the panel is not without cost either. In

fact, this increases the chances of attrition with every new wave, as well as the degree of computational difficulty in estimating qualitative limited dependent variable panel data models (see Baltagi, 2001).

Although RANDOM-COEFFICIENT MODELS can be used in the estimation and specification of panel data models (Hsiao, 1986), most panel data applications have been limited to a simple regression with error components disturbances, such as the following:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mu_i + v_{it}, \quad i = 1, \dots, N; t = 1, \dots, T$$

where y_{it} may denote log(wage) for the i th individual at time t , and $\mathbf{x}_{it}$ is a vector of observations on k explanatory variables such as education, experience, race, sex, marital status, union membership, hours worked, and so on. In addition, $\boldsymbol{\beta}$ is a k vector of unknown coefficients, μ_i is an unobserved individual specific effect, and v_{it} is a zero mean random disturbance with variance σ_v^2 . The error components disturbances follow a one-way analysis of variance (ANOVA). If μ_i denote fixed parameters to be estimated, this model is known as the FIXED-EFFECTS (FE) MODEL. The $\mathbf{x}_{it}$ s are assumed independent of the v_{it} s for all i and t . Inference in this case is conditional on the particular N individuals observed. Estimation in this case amounts to including $(N - 1)$ individual dummies to estimate these individual invariant effects. This leads to an enormous loss in degrees of freedom and attenuates the problem of MULTICOLLINEARITY among the regressors. Furthermore, this may not be computationally feasible for large N panels. In this case, one can eliminate the μ_i s and estimate $\boldsymbol{\beta}$ by running least squares of $\tilde{y}_{it} = y_{it} - \bar{y}_i$ on the $\tilde{\mathbf{x}}_{it}$ s similarly defined, where the dot indicates summation over that index and the bar denotes averaging. This transformation is known as the within transformation, and the corresponding estimator of $\boldsymbol{\beta}$ is called the within estimator or the FE estimator. Note that the FE estimator cannot estimate the effect of any time-invariant variable such as gender, race, religion, or union participation. These variables are wiped out by the within transformation. This is a major disadvantage if the effect of these variables on earnings is of interest. Ignoring the individual unobserved effects (i.e., running ordinary least squares [OLS] without individual dummies) leads to biased and inconsistent estimates of the regression coefficients.

If μ_i denotes independent random variables with zero mean and constant variance σ_μ^2 , this model is known as the random-effects model. The preceding moments are conditional on the $\mathbf{x}_{it}$ s. In addition, μ_i and v_{it} are assumed to be conditionally independent. The

random-effects (RE) model can be estimated by generalized least squares (GLS), which can be obtained using a least squares regression of $y_{it}^* = y_{it} - \theta \bar{y}_i$ on $\mathbf{x}_{it}^*$ similarly defined, where θ is a simple function of the variance components σ_μ^2 and σ_v^2 (Baltagi, 2001). The corresponding GLS estimator of $\boldsymbol{\beta}$ is known as the RE estimator. Note that for this RE model, one can estimate the effects of individual-invariant variables. The best quadratic unbiased (BQU) estimators of the variance components are ANOVA-type estimators based on the true disturbances, and these are minimum variance unbiased (MVU) under normality of the disturbances. One can obtain feasible estimates of the variance components by replacing the true disturbances by OLS or fixed-effects residuals. For the random-effects model, OLS is still unbiased and consistent but not efficient.

Fixed versus random effects has generated a lively debate in the biometrics and econometrics literature. In some applications, the random- and fixed-effects models yield different estimation results, especially if T is small and N is large. A specification test based on the difference between these estimates is given by Hausman (1978). The null hypothesis is that the individual effects are not correlated with the $\mathbf{x}_{it}$ s. The basic idea behind this test is that the fixed-effects estimator $\hat{\boldsymbol{\beta}}_{FE}$ is consistent, whether or not the effects are correlated with the $\mathbf{x}_{it}$ s. This is true because the fixed-effects transformation described by $\tilde{y}_{it}$ wipes out the μ_i effects from the model. In fact, this is the modern econometric interpretation of the FE model—namely, that the μ_i s are random but hopelessly correlated with all the $\mathbf{x}_{it}$ s. However, if the null hypothesis is true, the fixed-effects estimator is not efficient under the random-effects specification because it relies only on the within variation in the data. On the other hand, the random-effects estimator $\hat{\boldsymbol{\beta}}_{RE}$ is efficient under the null hypothesis but is biased and inconsistent when the effects are correlated with the $\mathbf{x}_{it}$ s. The difference between these estimators $\hat{q} = \hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE}$ tends to zero in probability limits under the null hypothesis and is nonzero under the alternative. The variance of this difference is equal to the difference in variances, $\text{var}(\hat{q}) = \text{var}(\hat{\boldsymbol{\beta}}_{FE}) - \text{var}(\hat{\boldsymbol{\beta}}_{RE})$ because $\text{cov}(\hat{q}, \hat{\boldsymbol{\beta}}_{RE}) = 0$ under the null hypothesis. Hausman's test statistic is based on $m = \hat{q}'[\text{var}(\hat{q})]^{-1}\hat{q}$ and is asymptotically distributed as a chi-square with k degrees of freedom under the null hypothesis.

For maximum likelihood as well as generalized method of moments estimation of panel models, the reader is referred to Baltagi (2001). Space limitations

do not allow discussion of panel data models that include treatment of missing observations, dynamics, measurement error, qualitative limited dependent variables, endogeneity, and nonstationarity of the regressors. Instead, we focus on some frequently encountered special panel data sets—namely, pseudo-panels and rotating panels. Pseudo-panels refer to the construction of a panel from repeated cross sections, especially in countries where panels do not exist but where independent surveys are available over time. The United Kingdom Family Expenditure Survey, for example, surveys about 7,000 households annually. These are independent surveys because it may be impossible to track the same household across surveys, as required in a genuine panel. Instead, one can track cohorts and estimate economic relationships based on cohort means. Pseudo-panels do not suffer the attrition problem that plagues genuine panels and may be available over longer time periods. One important question is the optimal size of the cohort. A large number of cohorts will reduce the size of a specific cohort and the samples drawn from it. Alternatively, selecting few cohorts increases the accuracy of the sample cohort means, but it also reduces the effective sample size of the panel.

Rotating panels attempt to keep the same number of households in the survey by replacing the fraction of households that drop from the sample in each period with an equal number of freshly surveyed households. This is a necessity in surveys in which a high rate of attrition is expected from one period to the next. Rotating panels allow the researcher to test for the existence of time-in-sample bias effects. These correspond to a significant change in response between the initial interview and a subsequent interview when one would expect the same response.

With the growing use of cross-country data over time to study purchasing power parity, growth convergence, and international research and development spillovers, the focus of panel data econometrics has shifted toward studying the asymptotics of macro panels with large N (number of countries) and large T (length of the time series) rather than the usual asymptotics of micro panels with large N and small T . Researchers argue that the time-series components of variables such as per capita gross domestic product growth have strong nonstationarity. Some of the distinctive results that are obtained with nonstationary panels are that many test statistics and estimators of interest have Normal limiting distributions. This is in

contrast to the nonstationary time-series literature in which the limiting distributions are complicated functionals of Weiner processes. Several unit root tests applied in the time-series literature have been extended to panel data (see Baltagi, 2001). However, the use of such panel data methods is not without their critics, who argue that panel data unit root tests do not rescue purchasing power parity (PPP). In fact, the results on PPP with panels are mixed depending on the group of countries studied, the period of study, and the type of unit root test used. More damaging is the argument that for PPP, panel data tests are the wrong answer to the low power of unit root tests in single time series. After all, the null hypothesis of a single unit root is different from the null hypothesis of a panel unit root for the PPP hypothesis. Similarly, panel unit root tests did not help settle the question of growth convergence among countries. However, it was useful in spurring much-needed research into dynamic panel data models.

Over the past 20 years, the panel data methodological literature has exhibited phenomenal growth. One cannot do justice to the many theoretical and empirical contributions to date. Space limitations prevented discussion of many worthy contributions. Some topics are still in their infancy but growing fast, such as nonstationary panels and semiparametric and nonparametric methods using panel data. It is hoped that this introduction will whet the reader's appetite and encourage more readings on the subject.

—Badi H. Baltagi

REFERENCES

- Baltagi, B. H. (2001). *Econometric analysis of panel data*. Chichester, UK: Wiley.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46, 1251–1271.
- Hsiao, C. (1986). *Analysis of panel data*. Cambridge, UK: Cambridge University Press.
- Kasprzyk, D., Duncan, G. J., Kalton, G., & Singh, M. P. (1989). *Panel surveys*. New York: John Wiley.

PARADIGM

In everyday usage, *paradigm* refers either to a model or an example to be followed or to an established system or way of doing things. The concept was introduced into the philosophy of science by Thomas

Kuhn (1970) in his discussion of the nature of scientific progress.

As a reaction against philosophies of science that prescribed *the* appropriate scientific method, such as Popper's FALSIFICATIONISM, Kuhn (1970) focused on the practices of communities of scientists. He saw such communities as sharing a paradigm or "discipline matrix" consisting of their views of the nature of the reality they study (their ONTOLOGY), including the components that make it up and how they are related; the techniques that are appropriate for investigating this reality (their EPISTEMOLOGY); and accepted examples of past scientific achievements (exemplars) that provide both the foundation for further practice and models for students who wish to become members of the community. He suggested that a mature science is dominated by a single paradigm.

According to Kuhn (1970), most of the time, scientists engage in *normal science*, research that is dominated by "puzzle-solving" activities and is firmly based on the assumptions and rules of the paradigm. Normal science extends the knowledge that the paradigm provides by testing its predictions and further articulating and filling out its implications; it does not aim for unexpected novelty of fact or theory. In the course of normal science, the paradigm is not challenged or tested; failure to solve a puzzle will be seen as the failure of the scientist, not the paradigm.

Occasions arise when some puzzles cannot be solved, or gaps appear between what the paradigm would have anticipated and what is observed. These anomalies may be ignored initially, as commitment to a paradigm produces inherent resistance to their recognition. Kuhn (1970) argued that a paradigm is a prerequisite to perception itself, that what we see depends both on what we look at and also on what our previous visual-conceptual experience has taught us to see. Adherence to a paradigm is analogous to an act of faith, and to suggest that there is something wrong with it is likely to be interpreted as heresy.

Anomalies may lead to a crisis of confidence in the paradigm. There emerges a period of *extraordinary science*, accompanied by a proliferation of competing articulations, the willingness to try anything, the expression of discontent, the recourse to philosophy, and debates over fundamentals. The situation is ripe for the emergence of a new paradigm and novel theories.

Kuhn (1970) has described the process of replacing the old paradigm with a new one as a *scientific*

revolution. A new paradigm may be proposed to replace an existing one—a paradigm that can solve the new puzzles raised by the anomalies and can handle the puzzles that the previous paradigm had solved. However, such revolutions occur only slowly, usually taking a generation or longer. According to Kuhn, the process by which a scientist moves from working with the old paradigm to the new is analogous to a religious conversion; it involves not just adopting a fundamentally different way of viewing the world but also living in a different world. Once a new paradigm is established, a new phase of normal science will commence. In time, new anomalies will emerge and further revolutions will occur.

Kuhn (1970) argued that rival paradigms are incommensurable. This is because the concepts and propositions of theories produced by a community of scientists depend on the assumptions and beliefs in their paradigm for their particular meaning. As paradigms embody different and incompatible world-views, it will be difficult for members of different scientific communities to communicate effectively, and it will be impossible to adjudicate between competing theories. There is no neutral language of observation, no common vocabulary, and no neutral ground from which to settle claims.

For Kuhn, scientific progress is not achieved by the accumulation of generalizations derived from observation (INDUCTION) or by the critical testing of new hypotheses (deduction/FALSIFICATIONISM)—it is achieved by scientific revolutions that change the way a scientific community views the world and defines and goes about solving puzzles (for reviews, see Blaikie, 1993, pp. 105–110; Riggs, 1992, pp. 22–59).

Kuhn's work has spawned a vast literature and has received detailed criticism by philosophers and historians of science, such as Lakatos and Laudan (for a review, see Riggs, 1992). It was used as a framework for understanding the crisis in sociology in the 1960s and 1970s (see, e.g., Friedrichs, 1970), when there were vigorous disputes between competing paradigms (see, e.g., Ritzer, 1980).

—Norman Blaikie

REFERENCES

- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Friedrichs, R. (1970). *A sociology of sociology*. New York: Free Press.

- Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.
- Riggs, P. J. (1992). *Whys and ways of science*. Melbourne, Australia: Melbourne University Press.
- Ritzer, G. (1980). *Sociology: A multiple paradigm science*. Boston: Allyn & Bacon.

PARAMETER

A parameter is any quantity that reveals a certain characteristic of a POPULATION. One conceptualization of a population is the set containing all outcomes of interest to the researcher. However, in statistics, a population is defined more technically as the set of all possible outcomes from a random EXPERIMENT. For example, the random experiment might consist of observations on the taxes paid by individual U.S. citizens. The population would then consist of the tax bill for each and every U.S. citizen. A population can be fully characterized by a RANDOM VARIABLE X (e.g., tax paid) and its associated PROBABILITY distribution $f(x)$ (e.g., a probability function or density function on tax paid). A population is considered known when we know the probability distribution $f(x)$ of the associated random variable X . If the population is known, parameters can be calculated easily from the population data. Parameters are therefore fixed quantities representing features of a population.

Two commonly used parameters in descriptive analysis are the MEAN and VARIANCE of a population. The population mean shows the numerical average of the population, whereas the population variance shows the average squared deviations of the observed values about the mean. For example, in the taxation experiment above, there is an average tax paid, as well as a variance in tax paid by all U.S. citizens. Other examples of parameters would be the population SKEWNESS, KURTOSIS, CORRELATION, and REGRESSION partial SLOPE parameter.

Parameters have no variability because they can be calculated from a known population. However, it is often inconvenient or impossible to work with a population. The population may be quite large, as in the taxpayer example above, and the researcher may not want to obtain or enter the data. It is also often difficult to obtain a population in the social sciences. The data may not be readily available to the researcher or may be nonexistent.

Most of the time in social science research, the population is not available, so the parameter is a theoretical construct that is unknown. The parameter exists somewhere in the real world, but researchers must estimate its value from a sample. In estimating a parameter from sample data, we regard all the entries of the sample as observations on *random variables* that are drawn from the population by random experiment. They are assumed to have the same probability distribution and to be drawn independently. *Sample statistics* that reveal certain characteristics of the sample are obtained from these random variables to estimate the population parameters. For example, in the tax experiment above, the sample mean is obtained by summing across all sample observations the products of sample entries and their respective probabilities. This is then considered as an estimator for population mean.

—B. Dan Wood and Sung Ho Park

REFERENCES

- Blalock, H. M., Jr. (1979). *Social statistics*. New York: McGraw-Hill.
- Casella, G., & Berger, R. L. (2001). *Statistical inference* (2nd ed.). New York: Wadsworth.
- Spiegel, M. R., Schiller, J., & Srinivasan, R. A. (2000). *Schaum's outline of probability and statistics* (2nd ed.). New York: McGraw-Hill.

PARAMETER ESTIMATION

Parameter estimation is concerned with finding the value of a population parameter from sample statistics. Population parameters are fixed and generally unknown quantities. If we know all possible entries of a population and their respective PROBABILITIES, we can calculate the value of a population parameter. However, we usually have only a sample from a population, so we have to estimate population parameters based on sample statistics. *Sample statistics* are constructed to reveal characteristics of the sample that may be considered ESTIMATORS of corresponding population parameters. Notice that sample statistics are themselves RANDOM VARIABLES because they are based on draws from a population.

Sample statistics are used as estimators of fixed population parameters. For example, the sample mean can be used as an estimator of the population mean. However, there is no guarantee that any particular

estimate, based on the estimator, will be an accurate reflection of the population parameter. There will always be uncertainty about the estimate due to SAMPLING ERROR. We can, however, minimize uncertainty about parameter estimates by using good estimators. A good estimator should satisfy certain statistical properties, such as UNBIASEDNESS, EFFICIENCY, asymptotic efficiency, or consistency. The first two are finite-sample properties, and the last two are asymptotic properties.

Unbiasedness requires that the expected value of an estimator should be equal to the true population parameter. That is, unbiasedness requires $E(\hat{\theta}) = \theta$, where θ is a population parameter and $\hat{\theta}$ is an estimator of the parameter. Notice again that this does not mean that a specific estimate will be equal to the true parameter. Instead, it is a property of repeated sampling. We expect only that the mean of the estimates that are calculated from repeated samples will be equal to the true parameter. *Efficiency* requires that the variance of an estimator be less than or equal to that of other unbiased estimators of the same parameter. That is, $\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2)$, where $\hat{\theta}_2$ comes from any estimator other than $\hat{\theta}_1$. It is sometimes the case that an estimator is biased but has smaller variance than other estimators that are unbiased. In this case, the minimum MEAN SQUARED ERROR (MSE) criterion can be useful. This is calculated as $\text{MSE}(\hat{\theta}|\theta) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta}|\theta)^2$. The estimator with the smallest MSE, regardless of bias, is generally considered the better estimator.

If an estimator satisfies both the unbiasedness and efficiency requirements, it is always a good estimator. However, it often happens that an estimator fails these requirements in small samples but, as the sample size increases, attains desirable statistical properties. Such estimators may be called asymptotically efficient or consistent. *Asymptotic efficiency* means that the variance of the asymptotic distribution of an estimator is smaller than the variance of any other asymptotically unbiased estimator. *Consistency* means that an estimator converges with unit probability to the value of the population parameter as we increase the sample size to infinity. These two properties are asymptotic equivalents of efficiency and unbiasedness in finite samples.

—B. Dan Wood and Sung Ho Park

REFERENCES

Blalock, H. M., Jr. (1979). *Social statistics*. New York: McGraw-Hill.

Casella, G., & Berger, R. L. (2001). *Statistical inference* (2nd ed.). New York: Wadsworth.

Judge, G. G., Hill, R. C., Griffiths, W. E., Lutkepohl, H., & Lee, T. (1988). *Introduction to the theory and practice of econometrics*. New York: John Wiley.

Spiegel, M. R., Schiller, J., & Srinivasan, R. A. (2000). *Schaum's outline of probability and statistics* (2nd ed.). New York: McGraw-Hill.

PART CORRELATION

The part correlation coefficient $r_{y(1.2)}$, which is also called the semipartial correlation coefficient by Hays (1972), differs from the partial correlation coefficient $r_{y1.2}$ in the following respect: The influence of the control variable X_2 is removed from X_1 but not from Y .

The three steps in PARTIAL CORRELATION analysis are now reduced to two. In the first step, the RESIDUAL scores $X_1 - \hat{X}_1$ from a bivariate REGRESSION analysis of X_1 on X_2 are calculated. In a second and last step, we calculate the ZERO-ORDER correlation coefficient between these residual scores $X_1 - \hat{X}_1$ and the original Y scores. This, then, is the part correlation coefficient $r_{y(1.2)}$. For HIGHER-ORDER coefficients, such as $r_{y(1.23)}$, it holds in the same way that the influence of the control variables X_2 and X_3 is removed from X_1 but not from Y . An example of $r_{y(1.2)}$ can be the correlation between the extracurricular activity of students (variable X_1) and their popularity (variable Y), controlling for their scholastic achievement (X_2). To clarify the mechanism of PART CORRELATION, we give a simple (fictitious) sample of five students with the following scores on the three variables: 1, 3, 2, 6, 4 for Y ; 0, 1, 3, 6, 8 for X_1 ; and 4, 4, 1, 2, 0 for X_2 . The part correlation coefficient $r_{y(1.2)}$ is now calculated as the ZERO-ORDER correlation between the $X_1 - \hat{X}_1$ scores -0.70 , 0.30 , -2.53 , 2.08 , 0.86 and the original Y scores 1, 3, 2, 6, 4, which yields $r_{y(1.2)} = 0.83$. This is then the correlation between popularity and extracurricular activity, in which the latter is cleared of the influence of scholastic achievement.

When dealing with considerations of content, part correlation finds little application. One reason why part correlation could be preferred to partial correlation in concrete empirical research is that the variance of Y would almost disappear when removing the influence of the control variable X_2 from it, so that there would be little left to explain. In such a case, the control variable would, of course, be a very special one because it

would share its variance almost entirely with one of the considered variables.

An example of a more appropriate use of the part correlation coefficient can be found in Colson's (1977, p. 216) research of the influence of the mean age (A) on the crude birth rate (B) in a sample of countries. But as the latter is not independent of the age structure and vice versa, because the mean age A contains a fertility component, a problem of contamination arose. This problem can be solved in two ways if we include the total (period) fertility rate as a variable, which is measured as the total of the age-specific fertility rates (F) and is independent of the age structure. A first procedure is the calculation of the correlation coefficient between A and F (i.e., the association between the mean age and age-free fertility). A second procedure, which was chosen by the author, is the calculation of the part correlation coefficient $r_{B(A.F)}$ between crude birth rates (measured as B) and the fertility-free age structure (residual scores of A after removing the influence of F).

This is an example of an appropriate use of part correlation analysis, provided that one is aware that one does not always know which concepts remain after the removal of contamination effects. It is therefore advisable, if possible, to preserve a conceptual distance between the original concepts (age structure and birth rate) to avoid contamination.

COMPUTATION OF THE PARTIAL AND PART CORRELATION

In the foregoing, we attempted to demonstrate the relevance of the part correlation in terms of contents. However, its power is predominantly formal in nature. Many formulas can be developed for the part correlation coefficient, and for each of these formulas, an equivalent for the partial correlation coefficient can be derived. We will now make some derivations for coefficient $r_{y(1.23)}$, in which original Y scores are correlated with scores of X_1 controlled for X_2 and X_3 .

The part correlation is, in its squared form, the increase in the proportion of the explained variance of Y that emerges when adding X_1 to the $\{X_2, X_3\}$ set:

$$r_{y(1.23)}^2 = R_{y.123}^2 - R_{y.23}^2.$$

We get a bit more insight into this formula when we rewrite it as follows:

$$R_{y.123}^2 = R_{y.23}^2 + r_{y(1.23)}^2.$$

The multiple determination coefficient $R_{y.123}^2$ is the proportion of the variance of Y that is explained by X_1 , X_2 , and X_3 together. This proportion can be split additively in two parts. The first part, $R_{y.23}^2$, is the proportion of the variance of Y that is explained by X_2 and X_3 together. The second part is the proportion that is also due to X_1 , after the latter is freed from the influence of X_2 and X_3 . This is the squared part correlation coefficient $r_{y(1.23)}^2$.

If X_1 is uncorrelated with X_2 and X_3 , then this increase in R^2 is simply equal to r_{y1}^2 . On the other hand, if X_1 is strongly associated with X_2 and X_3 , then the squared part correlation will be very different from the original squared correlation r_{y1}^2 .

Now, if the other variables X_2 and X_3 together explain a substantial proportion of the variance of Y , then there is not much left for X_1 to explain. In such a case, the part correlation $r_{y(1.23)}$ is unfairly small and will be hardly comparable with the situation in which the control variables do not explain a great amount. It is therefore wiser to divide the "absolute" increase in R^2 by the proportion of the variance of Y that is not explained by X_2 and X_3 . In doing so, we obtain the squared partial correlation coefficient, which is thus equal to the "relative" increase in the proportion of the explained variance of Y that emerges when adding X_1 to the $\{X_2, X_3\}$ set:

$$r_{y1.23}^2 = \frac{R_{y.123}^2 - R_{y.23}^2}{1 - R_{y.23}^2}.$$

Consequently, the squared partial correlation is obtained by dividing the squared part correlation by the part that is not explained by the control variables:

$$r_{y1.23}^2 = \frac{r_{y(1.23)}^2}{1 - R_{y.23}^2}$$

where

$$r_{y1.23} = \frac{r_{y(1.23)}}{(1 - R_{y.23}^2)^{\frac{1}{2}}}.$$

SIGNIFICANCE TEST

The significance tests of the part correlation coefficient and of the partial correlation coefficient are identical to the significance test of partial regression coefficients in regression analysis.

—Jacques Tacq

REFERENCES

- Blalock, H. M. (1972). *Social statistics*. Tokyo: McGraw-Hill Kogakusha.
- Brownlee, K. (1965). *Statistical theory and methodology in science and engineering*. New York: John Wiley.
- Colson, F. (1977). *Sociale indicatoren van enkele aspecten van bevolkingsgroei*. Doctoral dissertation, Katholieke Universiteit Leuven, Department of Sociology, Leuven.
- Hays, W. L. (1972). *Statistics for the social sciences*. New York: Holt.
- Tacq, J. J. A. (1997). *Multivariate analysis techniques in social science research: From problem to analysis*. London: Sage.

PARTIAL CORRELATION

Partial correlation is the correlation between two variables with the effect of (an)other variable(s) held constant. An often-quoted example is the correlation between the number of storks and the number of babies in a sample of regions, where the correlation disappears when the degree of rurality is held constant. This means that regions showing differences in the numbers of storks will also differ in birth rates, but this covariation is entirely due to differences in rurality. For regions with equal scores on rurality, this relationship between storks and babies should disappear. The correlation is said to be “spurious.”

As the correlation does not have to become zero but can change in whatever direction, we give here the example of the partial correlation between the extracurricular activity of students (variable X_1) and their popularity (variable Y), controlling for their scholastic achievement (X_2). To clarify the mechanism of partial correlation, we give a simple (fictitious) sample of five students with the following scores on the three variables:

Table 1

Y	X_1	X_2
1	0	4
3	1	4
2	3	1
6	6	2
4	8	0

To calculate the partial correlation, we remove the influence of X_2 from X_1 as well as from Y , so that the partial correlation becomes the simple ZERO-ORDER correlation between the residual terms $X_1 - \hat{X}_1$ and $Y - \hat{Y}$. This mechanism involves three steps.

First, a bivariate regression analysis is performed from X_1 on X_2 . The estimated $\hat{X}_1$ values are calculated, and the difference of $X_1 - \hat{X}_1$ yields the residual scores that are freed from control variable X_2 .

Second, a bivariate regression analysis of Y on X_2 is performed. The estimated $\hat{Y}$ values are calculated, and the difference of $Y - \hat{Y}$ yields the residual scores.

Third, $X_1 - \hat{X}_1$ and $Y - \hat{Y}$ are correlated. This zero-order correlation between the residuals is the partial correlation coefficient $r_{y1.2}$ (i.e., the zero-order product-moment correlation coefficient between Y and X_1 after the common variance with X_2 is removed from both variables).

This partial correlation coefficient is a FIRST-ORDER coefficient because one controls for only one variable. The model can be extended for HIGHER-ORDER partial correlations with several control variables. In the third-order coefficient $r_{y1.234}$, the partial correlation between Y and X_1 is calculated, controlling for X_2 , X_3 , and X_4 . In that case, the first step is not a bivariate but a multiple regression analysis of X_1 on X_2 , X_3 , and X_4 , resulting in the residual term $X_1 - \hat{X}_1$. The second step is a multiple regression analysis of Y on X_2 , X_3 , and X_4 , yielding the residual term $Y - \hat{Y}$. The zero-order correlation between $X_1 - \hat{X}_1$ and $Y - \hat{Y}$ is, then, the coefficient $r_{y1.234}$.

For the population, Greek instead of Latin letters are used. The sample coefficients $r_{y1.2}$ and $r_{y1.234}$, which are indicated above, are then written as $\rho_{y1.2}$ and $\rho_{y1.234}$.

CALCULATION OF THE PARTIAL CORRELATION

All the aforementioned operations are brought together in Table 2.

We can see here that the mechanism of partial correlation analysis should not be naively understood as the disappearance of a relationship when controlling for a third factor. In this case, the original correlation between Y and X_1 is $r_{y1} = 0.75$, and the partial correlation is $r_{y1.2} = 0.89$. So, the relationship between extracurricular activity and popularity appeared to be strong at first sight and now proves to be even stronger when controlling for scholastic achievement.

When calculating the partial correlation between Y and X_2 , controlling for X_1 (exercise for the reader), we clearly see how drastic the mechanism can be. The original correlation is $r_{y2} = -0.38$, and the partial correlation is $r_{y2.1} = 0.77$, which means that the sign has changed. The relationship between scholastic

Table 2

Step 1: $\hat{X}_1 = 7.14 - 1.61X_2$				Step 2: $\hat{Y} = 4.09 - 0.41X_2$			
X_1	X_2	$\hat{X}_1$	$X_1 - \hat{X}_1$	Y	X_2	$\hat{Y}$	$Y - \hat{Y}$
0	4	0.70	-0.70	1	4	2.45	-1.45
1	4	0.70	0.30	3	4	2.45	0.55
3	1	5.53	-2.53	2	1	3.68	-1.68
6	2	3.92	2.08	6	2	3.27	2.73
8	0	7.14	0.86	4	0	4.09	-0.09
Step 3:				$r_{(x_1 - \hat{x}_1)(y - \hat{y})} = r_{y 1.2} = 0.89$			

We see that the second-order partial correlation coefficient is a function of first-order coefficients. Each higher-order coefficient can be written in this way (i.e., in terms of coefficients of which the order is one lower).

SIGNIFICANCE TEST

The significance tests of the partial correlation coefficient and of the corresponding partial regression coefficient are identical. See REGRESSION ANALYSIS for further details.

—Jacques Tacq

REFERENCES

- Blalock, H. M. (1972). *Social statistics*. Tokyo: McGraw-Hill Kogakusha.
- Brownlee, K. (1965). *Statistical theory and methodology in science and engineering*. New York: John Wiley.
- Hays, W. L. (1972). *Statistics for the social sciences*. New York: Holt.
- Kendall, M., & Stuart, A. (1969). *The advanced theory of statistics: Vol. 2. Inference and relationship*. London: Griffin.
- Tacq, J. J. A. (1997). *Multivariate analysis techniques in social science research: From problem to analysis*. London: Sage.

PARTIAL LEAST SQUARES REGRESSION

Partial least squares (PLS) regression is a recent technique that generalizes and combines features from PRINCIPAL COMPONENTS ANALYSIS and MULTIPLE REGRESSION. It is particularly useful when we need to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors). It originated in the social sciences (specifically, economics; Wold, 1966) but became popular first in chemometrics, due in part to Wold's son Svante (see, e.g., Geladi & Kowalski, 1986), and in sensory evaluation (Martens & Naes, 1989). But PLS regression is also becoming a tool of choice in the social sciences as a multivariate technique for nonexperimental and experimental data alike (e.g., neuroimaging; see McIntosh, Bookstein, Haxby, & Grady, 1996). It was first presented as an algorithm akin to the power

method (used for computing eigenvectors) but was rapidly interpreted in a statistical framework (Frank & Friedman, 1993; Helland, 1990; Höskuldsson, 1988; Tenenhaus, 1998).

PREREQUISITE NOTIONS AND NOTATIONS

The I observations described by K dependent variables are stored in an $I \times K$ matrix denoted $\mathbf{Y}$, and the values of the J predictors collected on these I observations are collected in the $I \times J$ matrix $\mathbf{X}$.

GOAL

The goal of PLS regression is to predict $\mathbf{Y}$ from $\mathbf{X}$ and to describe their common structure. When $\mathbf{Y}$ is a vector and $\mathbf{X}$ is full rank, this goal could be accomplished using ORDINARY LEAST SQUARES (OLS). When the number of predictors is large compared to the number of observations, $\mathbf{X}$ is likely to be singular, and the regression approach is no longer feasible (i.e., because of MULTICOLLINEARITY). Several approaches have been developed to cope with this problem. One approach is to eliminate some predictors (e.g., using STEPWISE methods); another one, called principal component regression, is to perform a PRINCIPAL COMPONENTS ANALYSIS (PCA) of the $\mathbf{X}$ matrix and then use the principal components of $\mathbf{X}$ as regressors on $\mathbf{Y}$.

The orthogonality of the principal components eliminates the multicollinearity problem. But the problem of choosing an *optimum* subset of predictors remains. A possible strategy is to keep only a few of the first components. But they are chosen to explain $\mathbf{X}$ rather than $\mathbf{Y}$, and so nothing guarantees that the principal components, which “explain” $\mathbf{X}$, are relevant for $\mathbf{Y}$.

By contrast, PLS regression finds components from $\mathbf{X}$ that are also relevant for $\mathbf{Y}$. Specifically, PLS regression searches for a set of components (called *latent vectors*) that perform a simultaneous decomposition of $\mathbf{X}$ and $\mathbf{Y}$ with the constraint that these components explain as much as possible of the *covariance* between $\mathbf{X}$ and $\mathbf{Y}$. This step generalizes PCA. It is followed by a regression step in which the decomposition of $\mathbf{X}$ is used to predict $\mathbf{Y}$.

SIMULTANEOUS DECOMPOSITION OF PREDICTORS AND DEPENDENT VARIABLES

PLS regression decomposes both $\mathbf{X}$ and $\mathbf{Y}$ as a product of a common set of orthogonal factors and a set of specific loadings. So, the independent variables are *decomposed* as $\mathbf{X} = \mathbf{TP}^T$ with $\mathbf{T}^T\mathbf{T} = \mathbf{I}$, with $\mathbf{I}$ being the identity matrix (some variations of the technique do not require $\mathbf{T}$ to have unit norms). By analogy, with PCA, $\mathbf{T}$ is called the *score* matrix and $\mathbf{P}$ the *loading* matrix (in PLS regression, the loadings are not orthogonal). Likewise, $\mathbf{Y}$ is *estimated* as $\hat{\mathbf{Y}} = \mathbf{TBC}^T$, where $\mathbf{B}$ is a diagonal matrix with the “regression weights” as diagonal elements (see below for more details on these weights). The columns of $\mathbf{T}$ are the *latent vectors*. When their number is equal to the rank of $\mathbf{X}$, they perform an exact decomposition of $\mathbf{X}$. Note, however, that they only *estimate* $\mathbf{Y}$ (i.e., in general, $\hat{\mathbf{Y}}$ is not equal to $\mathbf{Y}$).

PLS REGRESSION AND COVARIANCE

The latent vectors could be chosen in a lot of different ways. In fact, in the previous formulation, any set of orthogonal vectors spanning the column space of $\mathbf{X}$ could be used to play the role of $\mathbf{Y}$. To specify $\mathbf{T}$, additional conditions are required. For PLS regression, this amounts to finding two sets of weights $\mathbf{w}$ and $\mathbf{c}$ to create (respectively) a linear combination of the columns of $\mathbf{X}$ and $\mathbf{Y}$ such that their covariance is maximum. Specifically, the goal is to obtain a first pair of vectors $\mathbf{t} = \mathbf{Xw}$ and $\mathbf{u} = \mathbf{Yc}$ with the constraints that $\mathbf{w}^T\mathbf{w} = 1$, $\mathbf{t}^T\mathbf{t} = 1$, and $\mathbf{t}^T\mathbf{u}$ be maximal. When the first latent vector is found, it is *subtracted* from both $\mathbf{X}$ and $\mathbf{Y}$ (this guarantees the orthogonality of the latent vectors), and the procedure is reiterated until $\mathbf{X}$ becomes a null matrix (see the algorithm section for more).

A PLS REGRESSION ALGORITHM

The properties of PLS regression can be analyzed from a sketch of the original algorithm. The first step is to create two matrices: $\mathbf{E} = \mathbf{X}$ and $\mathbf{F} = \mathbf{Y}$. These matrices are then column centered and normalized (i.e., transformed into Z -scores). The sums of squares of these matrices are denoted SS_X and SS_Y . Before starting the iteration process, the vector $\mathbf{u}$ is initialized with random values (in what follows, the symbol $\propto$ means “to normalize the result of the operation”).

- Step 1. $\mathbf{w} \propto \mathbf{E}^T\mathbf{u}$ (estimate $\mathbf{X}$ weights).
- Step 2. $\mathbf{t} \propto \mathbf{Ew}$ (estimate $\mathbf{X}$ factor scores).
- Step 3. $\mathbf{c} \propto \mathbf{F}^T\mathbf{t}$ (estimate $\mathbf{Y}$ weights).
- Step 4. $\mathbf{u} = \mathbf{Fc}$ (estimate $\mathbf{Y}$ scores).

If $\mathbf{t}$ has not converged, then go to Step 1; if $\mathbf{t}$ has converged, then compute the value of b , which is used to predict $\mathbf{Y}$ from $\mathbf{t}$ as $b = \mathbf{t}^T\mathbf{u}$, and compute the factor loadings for $\mathbf{X}$ as $\mathbf{p} = \mathbf{E}^T\mathbf{t}$. Now subtract (i.e., partial out) the effect of $\mathbf{t}$ from both $\mathbf{E}$ and $\mathbf{F}$ as follows $\mathbf{E} = \mathbf{E} - \mathbf{tp}^T$ and $\mathbf{F} = \mathbf{F} - b\mathbf{tc}^T$. The vectors $\mathbf{t}$, $\mathbf{u}$, $\mathbf{w}$, $\mathbf{c}$, and $\mathbf{p}$ are then stored in the corresponding matrices, and the scalar b is stored as a diagonal element of $\mathbf{B}$. The sum of squares of $\mathbf{X}$ (or, respectively, $\mathbf{Y}$) explained by the latent vector is computed as $\mathbf{p}^T\mathbf{p}$ (or, respectively, b^2), and the proportion of variance explained is obtained by dividing the explained sum of squares by the corresponding total sum of squares (i.e., SS_X and SS_Y).

If $\mathbf{E}$ is a null matrix, then the whole set of latent vectors has been found; otherwise, the procedure can be reiterated from Step 1 on.

PLS REGRESSION AND THE SINGULAR VALUE DECOMPOSITION

The iterative algorithm presented above is similar to the power method, which finds eigenvectors. So PLS regression is likely to be closely related to the eigenvalue and singular value decompositions, and this is indeed the case. For example, if we start from Step 1, which computes $\mathbf{w} \propto \mathbf{E}^T\mathbf{u}$, and substitute the rightmost term iteratively, we find the following series of equations: $\mathbf{w} \propto \mathbf{E}^T\mathbf{u} \propto \mathbf{E}^T\mathbf{Fc} \propto \mathbf{E}^T\mathbf{FF}^T\mathbf{t} \propto \mathbf{E}^T\mathbf{FF}^T\mathbf{Ew}$. This shows that the first weight vector $\mathbf{w}$ is the first right singular vector of the matrix $\mathbf{X}^T\mathbf{Y}$. Similarly, the first weight vector $\mathbf{c}$ is the left singular vector of $\mathbf{X}^T\mathbf{Y}$. The same argument shows that the first vectors $\mathbf{t}$ and $\mathbf{u}$ are the first eigenvectors of $\mathbf{XX}^T\mathbf{YY}^T$ and $\mathbf{YY}^T\mathbf{XX}^T$.

PREDICTION OF THE DEPENDENT VARIABLES

The dependent variables are predicted using the multivariate regression formula as $\hat{\mathbf{Y}} = \mathbf{TBC}^T = \mathbf{XB}_{\text{PLS}}$ with $\mathbf{B}_{\text{PLS}} = (\mathbf{P}^{T+})\mathbf{BC}^T$ (where $\mathbf{P}^{T+}$ is the Moore-Penrose pseudo-inverse of $\mathbf{P}^T$). If all the latent variables of $\mathbf{X}$ are used, this regression is equivalent to principal components regression. When only a subset of the latent variables is used, the prediction of $\mathbf{Y}$ is optimal for this number of predictors.

An obvious question is to find the number of latent variables needed to obtain the best generalization for the prediction of *new* observations. This is, in general, achieved by cross-validation techniques such as BOOTSTRAPPING.

The interpretation of the latent variables is often helped by examining graphs akin to PCA graphs (e.g., by plotting observations in a $t_1 \times t_2$ space).

A SMALL EXAMPLE

We want to predict the subjective evaluation of a set of five wines. The dependent variables that we want to predict for each wine are its likeability and how well it goes with meat or dessert (as rated by a panel of experts) (see Table 1). The predictors are the price and the sugar, alcohol, and acidity content of each wine (see Table 2).

Table 2 The Y Matrix of Dependent Variables

<i>Wine</i>	<i>Hedonic</i>	<i>Goes With Meat</i>	<i>Goes With Dessert</i>
1	14	7	8
2	10	7	6
3	8	5	5
4	2	4	7
5	6	2	4

Table 3 The X Matrix of Predictors

<i>Wine</i>	<i>Price</i>	<i>Sugar</i>	<i>Alcohol</i>	<i>Acidity</i>
1	7	7	13	7
2	4	3	14	7
3	10	5	12	5
4	16	7	11	3
5	13	3	10	3

The different matrices created by PLS regression are given in Tables 3 to 11. From Table 11, one can find that two latent vectors explain 98% of the variance of **X** and 85% of **Y**. This suggests keeping these two dimensions for the final solution. The examination of the two-dimensional regression coefficients (i.e., **B**_{PLS}; see Table 8) shows that sugar content is mainly responsible for choosing a dessert wine and that price is negatively correlated with the perceived quality of the wine, whereas alcohol content is positively correlated with it (at least in this example). The latent vectors show that t_1 expresses price and t_2 reflects sugar content.

Table 4 The Matrix **T**

<i>Wine</i>	t_1	t_2	t_3
1	0.4538	−0.4662	0.5716
2	0.5399	0.4940	−0.4631
3	0	0	0
4	−0.4304	−0.5327	−0.5301
5	−0.5633	0.5049	0.4217

Table 5 The Matrix **U**

	u_1	u_2	u_3
1	1.9451	−0.7611	0.6191
2	0.9347	0.5305	−0.5388
3	−0.2327	0.6084	0.0823
4	−0.9158	−1.1575	−0.6139
5	−1.7313	0.7797	0.4513

Table 6 The Matrix **P**

	p_1	p_2	p_3
Price	−1.8706	−0.6845	−0.1796
Sugar	0.0468	−1.9977	0.0829
Alcohol	1.9547	0.0283	−0.4224
Acidity	1.9874	0.0556	0.2170

Table 7 The Matrix **W**

	w_1	w_2	w_3
Price	−0.5137	−0.3379	−0.3492
Sugar	0.2010	−0.9400	0.1612
Alcohol	0.5705	−0.0188	−0.8211
Acidity	0.6085	0.0429	0.4218

Table 7 The Matrix **B**_{PLS} When Three Latent Vectors Are Used

	<i>Hedonic</i>	<i>Goes With Meat</i>	<i>Goes With Dessert</i>
Price	−1.0607	−0.0745	0.1250
Sugar	0.3354	0.2593	0.7510
Alcohol	−1.4142	0.7454	0.5000
Acidity	1.2298	0.1650	0.1186

RELATIONSHIP WITH OTHER TECHNIQUES

PLS regression is obviously related to CANONICAL CORRELATION and to multiple FACTOR ANALYSIS (Escofier & Pagès, 1988). These relationships are explored in detail by Tenenhaus (1998) and Pagès and Tenenhaus (2001). The main originality of PLS regression is preserving the asymmetry of the relationship

Table 8 The Matrix $\mathbf{B}_{\text{PLS}}$ When Two Latent Vectors Are Used

	<i>Hedonic</i>	<i>Goes With Meat</i>	<i>Goes With Dessert</i>
Price	−0.2662	−0.2498	0.0121
Sugar	0.0616	0.3197	0.7900
Alcohol	0.2969	0.3679	0.2568
Acidity	0.3011	0.3699	0.2506

Table 9 The Matrix $\mathbf{C}$

	$\mathbf{c}_1$	$\mathbf{c}_2$	$\mathbf{c}_3$
Hedonic	0.6093	0.0518	0.9672
Goes with meat	0.7024	−0.2684	−0.2181
Goes with dessert	0.3680	−0.9619	−0.1301

Table 10 The $\mathbf{b}$ Vector

b_1	b_2	b_3
2.7568	1.6272	1.1191

Table 11 Variance of $\mathbf{X}$ and $\mathbf{Y}$ Explained by the Latent Vectors

<i>Latent Vector</i>	<i>% of Explained Variance for $\mathbf{X}$</i>	<i>Cumulative % of Explained Variance for $\mathbf{X}$</i>	<i>% of Explained Variance for $\mathbf{Y}$</i>	<i>Cumulative % of Explained Variance for $\mathbf{Y}$</i>
1	70	70	63	63
2	28	98	22	85
3	2	100	10	95

between predictors and dependent variables, whereas these other techniques treat them symmetrically.

SOFTWARE

PLS regression necessitates sophisticated computations, and its application depends on the availability of software. For chemistry, two main programs are used: the first one, called SIMCA-P, was originally developed by Wold; the second one, called the UNSCRAMBLER, was first developed by Martens, who was another pioneer in the field. For brain imaging, SPM, which is one of the most widely used programs in the field, has recently (2002) integrated a PLS regression module. Outside these fields, SAS PROC PLS is probably the most easily available program. In addition, interested readers can download a set of MATLAB programs from the author's home page (www.utdallas.edu/~herve). A public domain set of MATLAB programs

is available from the home page of the *N*-Way project (www.models.kvl.dk/source/nwaytoolbox/), along with tutorials and examples. Finally, a commercial MATLAB toolbox has also been developed by EIGENRESEARCH.

—Hervé Abdi

REFERENCES

- Escofier, B., & Pagès, J. (1988). *Analyses factorielles multiples* [Multiple factor analyses]. Paris: Dunod.
- Frank, I. E., & Friedman, J. H. (1993). A statistical view of chemometrics regression tools. *Technometrics*, 35, 109–148.
- Geladi, P., & Kowalski, B. (1986). Partial least square regression: A tutorial. *Analytica Chimica Acta*, 35, 1–17.
- Helland, I. S. (1990). PLS regression and statistical models. *Scandinavian Journal of Statistics*, 17, 97–114.
- Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, 2, 211–228.
- Martens, H., & Naes, T. (1989). *Multivariate calibration*. London: Wiley.
- McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. (1996). Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage*, 3, 143–157.
- Pagès, J., & Tenenhaus, M. (2001). Multiple factor analysis combined with PLS path modeling: Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgments. *Chemometrics and Intelligent Laboratory Systems*, 58, 261–273.
- Tenenhaus, M. (1998). *La régression PLS* [PLS regression]. Paris: Technip.
- Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate analysis* (pp. 391–420). New York: Academic Press.

PARTIAL REGRESSION COEFFICIENT

The partial regression coefficient is also called the REGRESSION COEFFICIENT, regression weight, partial regression weight, SLOPE coefficient, or partial slope coefficient. It is used in the context of multiple linear regression (MLR) analysis and gives the amount by which the DEPENDENT VARIABLE (DV) increases when one INDEPENDENT VARIABLE (IV) is increased by one unit and all the other independent variables are held constant. This coefficient is called partial because its value depends, in general, on the other independent variables. Specifically, the value of the partial coefficient for one independent variable will vary, in general,

depending on the other independent variables included in the regression equation.

MULTIPLE REGRESSION FRAMEWORK

In MLR, the goal is to PREDICT, knowing, from the measurements collected on N subjects, the value of the dependent variable Y from a set of K independent variables $\{X_1, \dots, X_k, \dots, X_K\}$. We denote by $\mathbf{X}$ the $N \times (K + 1)$ augmented MATRIX collecting the data for the independent variables (this matrix is called *augmented* because the first column is composed only of 1s), and we denote by $\mathbf{y}$ the $N \times 1$ vector of observations for the dependent variable (see Figure 1).

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} & \cdots & x_{1K} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} & \cdots & x_{nK} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nk} & \cdots & x_{NK} \end{pmatrix}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \\ y_N \end{pmatrix}.$$

Figure 1 The Structure of the X and y Matrices

Multiple regression finds a set of partial regression coefficients b_k such that the dependent variable could be approximated as well as possible by a linear combination of the independent variables (with the b_k s being the weights of the combination). Therefore, a predicted value, denoted $\hat{Y}$, of the dependent variable is obtained as

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + \cdots + b_K X_K. \quad (1)$$

The value of the partial coefficients is found using ORDINARY LEAST SQUARES (OLS). It is often convenient to express the MLR equation using matrix notation. In this framework, the predicted values of the dependent variable are collected in a vector denoted $\hat{\mathbf{y}}$ and are obtained using MLR as

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \quad \text{with} \quad \mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (2)$$

The quality of the prediction is evaluated by computing the multiple coefficient of correlation squared $R_{Y,1,\dots,k,\dots,K}^2$, which is the squared coefficient of correlation between the dependent variable (Y) and the predicted dependent variable ($\hat{Y}$). The *specific* contribution of each IV to the regression equation is assessed by the partial coefficient of correlation associated with each variable. This coefficient, closely associated with the partial regression coefficient, corresponds to the increment in explained variance obtained by adding this variable to the regression equation after all the other IVs have been already included.

PARTIAL REGRESSION COEFFICIENT AND REGRESSION COEFFICIENT

When the independent variables are pairwise orthogonal, the effect of each of them in the regression is assessed by computing the slope of the regression between this independent variable and the dependent variable. In this case (i.e., orthogonality of the IVs), the partial regression coefficients are equal to the regression coefficients. In all other cases, the regression coefficient will differ from the partial regression coefficients.

Table 1 A Set of Data: Y Is Predicted From X_1 and X_2

Y (Memory Span)	14	23	30	50	39	67
X_1 (Age)	4	4	7	7	10	10
X_2 (Speech Rate)	1	2	2	4	3	6

SOURCE: Abdi, Dowling, Valentin, Edelman, and Posamentier (2002). NOTE: Y is the number of digits a child can remember for a short time (the "memory span"), X_1 is the age of the child, and X_2 is the speech rate of the child (how many words the child can pronounce in a given time). Six children were tested.

For example, consider the data given in Table 1, in which the dependent variable Y is to be predicted from the independent variables X_1 and X_2 . In this example, Y is a child's memory span (i.e., number of words a child can remember in a set of short-term memory tasks) that we want to predict from the child's age (X_1) and the child's speech rate (X_2). The prediction equation (using equation (2)) is $\hat{Y} = 1.67 + X_1 + 9.50X_2$, where b_1 is equal to 1 and b_2 is equal to 9.50. This means that children increase their memory span by one word every year and by 9.50 words for every additional word they can pronounce (i.e., the faster they speak, the more they can remember).

A multiple linear regression analysis of this data set gives a multiple coefficient of correlation squared of $R^2_{Y \cdot 1,2} = .9866$. The coefficient of correlation between X_1 and X_2 is equal to $r_{1,2} = .7500$, between X_1 and Y is equal to $r_{Y,1} = .8028$, and between X_2 and Y is equal to $r_{Y,2} = .9890$. The squared partial regression coefficient between X_1 and Y is computed as $r^2_{Y \cdot 1|2} = R^2_{Y \cdot 1,2} - r^2_{Y,2} = .9866 - .9890^2 = .0085$; likewise, $r^2_{Y \cdot 2|1} = R^2_{Y \cdot 1,2} - r^2_{Y,1} = .9866 - .8028^2 = .3421$.

F AND T-TESTS FOR THE PARTIAL REGRESSION COEFFICIENT

The NULL HYPOTHESIS stating that a partial regression coefficient is equal to zero can be tested by using a standard F TEST, which tests the equivalent null hypothesis stating that the associated partial coefficient of correlation is zero. This F test has $\nu_1 = 1$ and $\nu_2 = N - K - 1$ degrees of freedom (with N being the number of observations and K being the number of predictors). Because ν_1 is equal to 1, the square root of F gives a Student t -test. The computation of F is best described with an example: The F for the variable X_1 in our example is obtained as follows:

$$\begin{aligned} F_{Y \cdot 1,2} &= \frac{r^2_{Y \cdot 1|2}}{1 - R^2_{Y \cdot 1,2}} \times df_{\text{regression}} \\ &= \frac{r^2_{Y \cdot 1|2}}{1 - R^2_{Y \cdot 1,2}} \times (N - K - 1) \\ &= \frac{.0085}{1 - .9866} \times 3 = 1.91. \end{aligned}$$

This value of F is smaller than the critical value for $\nu_1 = 1$ and $\nu_2 = N - K - 1 = 3$, which is equal to 10.13 for an alpha level of .05. Therefore, b_1 (and $r^2_{Y \cdot 1|2}$) cannot be considered different from zero.

STANDARD ERROR AND CONFIDENCE INTERVAL

The STANDARD ERROR of the partial regression coefficient is useful to compute CONFIDENCE INTERVALS and perform additional statistical tests. The standard error of coefficient b_k is denoted S_{b_k} . It can be computed directly from F as $S_{b_k} = b_k \times \sqrt{1/F_k}$, where F_k is the value of the F test for b_k . For example, we find for the first variable that: $S_{b_1} = 1 \times \sqrt{1/1.91} = .72$.

The confidence interval of b_j is computed as $b_k \pm S_{b_k} \sqrt{F_{\text{critical}}}$, with F_{critical} being the critical value for

the F test. For example, the confidence interval of b_1 is equal to $1 \pm .72 \times \sqrt{10.13} = 1 \pm 2.29$, and therefore the 95% confidence interval b_1 goes from -1.29 to $+3.29$. This interval encompasses the zero value, and this corresponds to the failure to reject the null hypothesis.

β WEIGHTS AND PARTIAL REGRESSION COEFFICIENTS

There is some confusion here because, depending on the context, β is the parameter estimated by b , whereas in some other contexts, β is the regression weight obtained when the regression is performed with all variables being expressed in Z -scores. In the latter case, β is called the *standardized partial regression coefficient* or the β -weight. These weights have the advantage of being comparable from one independent variable to the other because the unit of measurement has been eliminated. The β -weights can easily be computed from the b s as $\beta_k = \frac{S_k}{S_y} b_k$ (with S_k being the standard deviation of the k th independent variable and S_y being the standard deviation of the dependent variable).

—Hervé Abdi

REFERENCES

- Abdi, H., Dowling, W. J., Valentin, D., Edelman, B., & Posamentier, M. (2002). *Experimental design and research methods*. Unpublished manuscript, University of Texas at Dallas, Program in Cognition and Neuroscience.
- Cohen, J., & Cohen, P. (1984). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.
- Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. New York: John Wiley.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). New York: Wadsworth.

PARTICIPANT OBSERVATION

Participant observation is a method of data collection in which the investigator uses participation in an area of ongoing social life to observe it. The investigator may or may not already be a natural member of the social setting studied. The classic form that such participation takes has been relatively

long-term membership in a small face-to-face group, but it can also extend to the anonymous joining of a crowd; the boundary between participant and nonparticipant observation is thus not a clear one. Raymond L. Gold's (1969) distinctions among possible field roles—from complete participant to participant as observer/observer as participant/complete observer—have been widely used to indicate the range of possibilities. What participant observers actually do can vary considerably and may include elements sometimes counted as separate methods in their own right such as UNSTRUCTURED INTERVIEWS, observation of physical features of settings, or the collection of printed materials.

It has become conventional to classify participant observation as a qualitative method, but its data can sometimes be quantified, even if the circumstances may often make the collection of standardized material difficult. Equally, it is conventionally assumed that the observer does not set out to test hypotheses but enters the field with an open mind to see what life is like. The mode of data collection as such does not prevent the testing of hypotheses, but the assumption is that adequate hypotheses cannot be formed without close familiarity with a situation and the meanings of members in it. Participant observation has mainly been used for studies of groups in work settings or of informal groupings in small communities or the social life of groups of friends; these are situations in which patterns of face-to-face interaction develop, and a single researcher or a small group of workers can gain access and establish a role. However, there also, for example, have been studies of football crowds, right-wing political groups, and schools. The method is not well suited to the study of large populations or to the collection of representative data about causal relations among predefined variables; the logic of the CASE STUDY rather than of the RANDOM SAMPLE applies.

HISTORY

The history of the *practice* of “participant observation” needs to be distinguished from the history of the use of the *term* and of its application in the modern sense. Observers reported on their findings from personal participation before the term emerged, some well before the development of social science as an academic field. The term was first used in the literature of sociology, in 1924, for the role of informants

who reported to the investigator on meetings they had attended. (Bronislaw Malinowski was inventing the form that it came to take in social-anthropological fieldwork in the Trobriand islands during World War I, although the anthropological and sociological traditions were largely independent.) The body of work produced at the University of Chicago in the 1920s and 1930s, under the leadership of Robert Park, is generally regarded as providing the first examples in sociology of participant observation in the modern sense, although they did not then call it that. Some of these studies were ones in which Park encouraged his students to write up prior experiences, or ones that arose naturally in their daily lives, rather than undertaking them for research purposes. These studies drew on many sources of data, without privileging participation as giving special access to meanings. It was only after World War II that the methodological literature developed, again led by Chicago and authors there such as Howard S. Becker, which focused on participant observation as one of the major alternative modes of data collection. The direct empathetic access it provided to real social behavior and its meanings was contrasted with the newly dominant questionnaire survey, and its data were seen as superior to the answers to standardized questions that merely reported on behavior.

PRACTICAL AND ETHICAL PROBLEMS

Some special research problems, practical and ethical, follow from the use of participant observation. ACCESS to the research situation needs to be gained. Negotiation with gatekeepers, when identified, may solve that for overt research. But what particular researchers can do depends, more than with any other method of data collection, on their personal characteristics; a man and a woman, or a Black and a White researcher, do not have the same ranges of opportunities, although such conventional boundaries have sometimes been transcended. Conducting research covertly may solve some problems of access but, in addition to raising acute ethical issues, makes the possibility of exposure a continual anxiety; important work, however, has been done covertly that could not have been done overtly. Even when participant observation is overt, research subjects can forget, once the researcher has become a friend, that their interaction may become data; the researcher's dilemma is then that continual renewal of consent would jeopardize

the naturalness of the relationship and so damage the data. This is part of two broader problems: how to live psychologically with using the self for research purposes, perhaps for many hours a day, without GOING NATIVE, and how to deal with the fact that the researcher's presence must, to some extent, change the situation that is being studied. The issue of sampling as such is seldom addressed for participant observation but arises in the form of the need to avoid relying on unrepresentative informants and to observe interaction at different times and places in the life of the group studied. The physical task of recording the data has to be fitted in around making the observations; solutions have ranged from writing up notes every night from memory to frequent trips to the toilet or the use of pocket counting devices. How satisfactory such arrangements are will depend on the social situation and the nature of the data sought.

EXAMPLES

Three well-known examples of the method are William F. Whyte's (1943/1955) *Street Corner Society*; Howard S. Becker, Everett C. Hughes, and Anselm L. Strauss's (1961) *Boys in White*; and Laud Humphreys's (1970) *Tearoom Trade*. Whyte's study of a gang in Boston is not only of substantive interest but also important for the detailed methodological appendix added to later editions, which did much to create discussion of the practical issues involved in such work. *Boys in White* was a study of medical students that adopted exceptionally systematic and clearly documented methods of observation and analysis. *Tearoom Trade* has become notorious for its observation of homosexual activity in public toilets, as well as the ethical issues raised by Humphreys's indirect methods of identifying the participants.

—Jennifer Platt

REFERENCES

- Becker, H. S., Hughes, E. C., & Strauss, A. L. (1961). *Boys in white*. Chicago: University of Chicago Press.
- Gold, R. L. (1969). Roles in sociological field observations. In G. J. McCall & J. L. Simmons (Eds.), *Issues in participant observation* (pp. 30–38). Reading, MA: Addison-Wesley.
- Humphreys, L. (1970). *Tearoom trade*. Chicago: Aldine.
- Whyte, W. F. (1955). *Street corner society*. Chicago: University of Chicago Press. (Original work published in 1943.)

PARTICIPATORY ACTION RESEARCH

Participatory action research (PAR) is research involving the collaboration between local communities, organizations, or coalitions with a legitimate personal interest in solving a problem that affects them directly and expert outside facilitators. Together they define the problem, learn how to study it, design the research, analyze the outcomes, and design and execute the needed actions. PAR rests on a belief in the importance of local knowledge and the ability of people of all social conditions to take action on their own behalf to democratize the situations they live in.

Key elements in participatory action research include the belief that the purpose of research is to promote democratic social change by enhancing the capacity of people to chart the course of their own future, that insider-outsider collaboration is both possible and desirable, that local knowledge is essential to the success of any social change project, and that nothing of importance happens if power imbalances are not ameliorated to some degree. PAR also aims to enhance local stakeholders' ability to take on further change projects on their own rather than relying on outside experts.

A wide variety of ideological positions and methodological practices are associated with PAR. Some practitioners believe in a strong model of external leadership or even provocation. Others believe that people can organize themselves effectively if given a favorable protected space to work in. Some see radical social change as the defining characteristic of PAR, whereas others believe that any changes that enhance well-being, sustainability, or fairness are sufficient.

The terminological field surrounding PAR is dense and complex. Many varieties of activist research bear some relationship to each other (see the entry on ACTION RESEARCH for additional information). Among these are action research, action science, action learning, collaborative inquiry, and the methodological traditions of GROUNDED THEORY and general systems theory. These approaches differ in the disciplinary backgrounds they draw on (education, sociology, social movement analysis, applied anthropology, psychology, social work). They differ in what they define as a good outcome, and they differ in how strong a role expert outsiders are expected to play in the change process.

PAR has long had a controversial relationship to the academic disciplines and to the academy in general. A purposely engaged form of research, PAR rejects the separation of thought and action on which the academic social sciences and humanities generally rest and fundamentally questions the notions of objectivity, control, distance, and quality control by disciplinary standards.

PAR is practiced in many different contexts—communities, organizations, a business unit, a voluntary coalition, and so on—and so no single example can do justice to the variety of approaches. There are diverse ways of beginning a project: being invited as an outside consultant to a group or coalition, being a member of a thematic organization known for expertise with certain issues, being invited to work with community partners, and so forth. How one begins matters, but what matters most is not how one begins but the democratic quality of the process as it develops—its openness, its flexibility, its consistency with democratic purposes. Despite all the differences, a single brief example, drawn from my own work, will help clarify how such projects begin and develop.

In a small town in south central Spain, the agrarian economy has become increasingly vulnerable. More than 300 men leave daily to work in construction in Madrid, and hundreds of women do piecework sewing. A highly stratified community with a history of class conflict, the town faces the loss of most of its young people to other cities, something affecting all classes equally in a society that strongly values family ties. To address this, we convened a planning group of municipal officials and public and private school teachers who planned an action research search conference for 40 people selected from all ages, classes, genders, and ideologies. This 2-day event involved developing a shared view of the history, the ideal future, and the probable future of the community; the development of plans to improve future prospects for the young; and the creation of voluntary action teams to work on the different problems identified. The teams met for months after this, elaborating plans, acquiring resources, and taking actions with the outside facilitator in contact by e-mail and reconvening groups at 3-month intervals in person.

The concrete experience of the ability of people with very different social positions, wealth, and ideologies to collaborate and discover that each held a “piece of the puzzle” was a lasting contribution of the project. A number of the small action projects prospered, but

some also failed. Now, elements and processes drawn from those initial activities are coalescing in a public initiative to create an ethnographic museum, civic center, and in tourism promotion efforts.

—Davydd J. Greenwood

REFERENCES

- Fals-Borda, O., & Anisur Rahman, M. (Eds.). (1991). *Action and knowledge: Breaking the monopoly with participatory action-research*. New York: Apex.
- Greenwood, D., & Levin, M. (1998). *Introduction to action research: Social research for social change*. Thousand Oaks, CA: Sage.
- Horton, M., & Freire, P. (2000). *We make the road by walking* (Bell, B., Gaventa, J., & Peters, J., Eds.). Philadelphia: Temple University Press.
- Reason, R., & Bradbury, H. (Eds.). (2001). *The handbook of action research*. London: Sage.
- Whyte, W. F. (Ed.). (1991). *Participatory action research*. Newbury Park, CA: Sage.

PARTICIPATORY EVALUATION

Participatory evaluation is an umbrella term for approaches to program evaluation that actively involve program staff and participants in the planning and implementation of evaluation studies. This involvement ranges on a continuum from shared responsibility for evaluation questions and activities, including data collection and analysis, to complete control of the evaluation process and its outcomes, where the evaluator serves as a coach to ensure the quality of the effort. *Collaborative evaluation*, a related term, is a form of participatory evaluation near the center of this continuum; program participants take an active role, for example, in designing questions and data collection instruments, but the program evaluator remains a partner in the process, rather than a coach.

What all forms of participatory evaluation share is the purposeful and explicit involvement of program participants in the evaluation process to effect change, either in a program or organization locally or in society more generally. Many also seek to build the capacity of an organization or group to conduct additional evaluation without the assistance of a professional evaluator. Weaver and Cousins (2001) provide five dimensions for analyzing participatory evaluation studies: the extent to which stakeholders control decisions about

the evaluation, the diversity of individuals selected to participate in the discussions, the power relations of participants, the ease with which the study can be implemented, and the depth of people's participation.

The origins of participatory evaluation reflect the general nature of the term, with roots in disparate traditions from around the world. Building on the work of Cousins and Whitmore (1998), theorists distinguish between two categories of approaches: practical participatory evaluation (P-PE) and transformative participatory evaluation (T-PE). Practical participatory evaluation evolved from approaches that seek to increase the ownership and use of evaluation results by involving program staff and participants in the evaluation process. Michael Quinn Patton's (1997) utilization-focused evaluation highlights the importance of the "personal factor" (i.e., connecting with individuals who are interested in the evaluation and providing information of use to them). Organization development and organizational learning point to the value of cyclic processes over time for asking questions, collecting and analyzing data, reflecting on them, and making changes as appropriate. In P-PE, evaluators can provide leadership of the evaluation process as they teach it to others by having them engage in it.

In T-PE, by contrast, program participants (at least theoretically) become the leaders of evaluation efforts to support social change and social justice through the development of empowered citizens. Whereas P-PE can improve the status quo by making programs more effective or efficient (e.g., improving service delivery in mental health programs), T-PE is a process that seeks to better the world for disenfranchised and oppressed people (e.g., reducing poverty among rural communities in a developing country). This category now encompasses a number of citizen-based evaluation/research approaches of the past 50 years, including, for example, participatory action research, participatory rural appraisal (renamed *participatory learning and action*), participatory monitoring and evaluation, and, more recently, empowerment evaluation.

—Jean A. King

REFERENCES

- Cousins, J. B. (2003). Utilization effects of participatory evaluation. In T. Kellaghan, D. L. Stufflebeam, & L. A. Wingate (Eds.), *International handbook of educational evaluation* (pp. 245–266). Dordrecht, The Netherlands: Kluwer Academic.
- Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. *New Directions in Evaluation*, 80, 5–23.
- King, J. A. (1998). Making sense of participatory evaluation practice. *New Directions for Evaluation*, 80, 57–67.
- Patton, M. Q. (1997). *Utilization-focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage.
- Weaver, L., & Cousins, B. (2001, November). *Unpacking the participatory process*. Paper presented at the annual meeting of the American Evaluation Association, St. Louis, MO.

PASCAL DISTRIBUTION

The Pascal distribution is used to model the number of independent BERNOULLI trials (x) needed to obtain a specific number of successes (r). The discrete RANDOM VARIABLE X can take on any integer value from r to infinity. If one is lucky and every trial results in a success, then $X = r$. To have $X = x$, it is necessary for there to be a success in trial x following exactly $r - 1$ successes during the preceding $x - 1$ trials. The definition here is opposite to that of the BINOMIAL distribution. That is, the random variable in the Pascal distribution is the number of trials before a certain number of successes, whereas in the binomial distribution, the random variable is the number of successes within a certain number of trials. For this reason, the Pascal distribution is also referred to as the NEGATIVE BINOMIAL DISTRIBUTION. Although these two distributions are often used synonymously, the Pascal distribution is actually a special case of the negative binomial in which the random variable X must be an integer. A special case of the Pascal distribution, the geometric distribution, occurs when $r = 1$, that is, how many trials are necessary to obtain the first success.

The probability mass function (pmf) for the Pascal distribution is given by

$$P(X = x|r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

for $x = r, r + 1, r + 2, \dots$,

$$0 \leq p \leq 1, 0 < r \leq \infty.$$

The mean of a Pascal distributed random variable is

$$\mu = \frac{r}{p}$$

and the variance is

$$\sigma^2 = \frac{r(1-p)}{p^2}.$$

The Pascal distribution is an extension of the POISSON DISTRIBUTION in which the restrictions on the parameter λ are relaxed. Rather than requiring a constant rate of events over an observation period, the Pascal distribution assumes this rate to vary according to the gamma distribution. For this reason, the Pascal distribution is more flexible than the Poisson distribution and is often used to model overdispersed EVENT COUNT data, or data in which the variance exceeds the mean.

For example, the Pascal distribution could be used to model the number of bills that must be proposed to a legislature before 10 bills are passed. Alternatively, the results of such an experiment can be used to determine the number of failures before the r th success.

—Ryan Bakker

REFERENCES

- Johnson, N. L., & Kotz, S. (1969). *Discrete distributions*. Boston: Houghton-Mifflin.
- King, G. (1998). *Unifying political methodology: The likelihood theory of statistical inference*. Ann Arbor: University of Michigan Press.

PATH ANALYSIS

Although the dictum “Correlation does not imply causation” is well rehearsed in research methods courses throughout the social sciences, its converse actually serves as the basis for *path analysis* and the larger STRUCTURAL EQUATION MODELING (SEM) family of quantitative data analysis methods (which also includes, for example, MULTIPLE REGRESSION ANALYSIS and CONFIRMATORY FACTOR ANALYSIS). Specifically, under the right circumstances, “Causation does imply correlation.” That is to say, if one variable has a causal bearing on another, then a CORRELATION (or COVARIANCE) should be observed in the data when all other relevant variables are controlled statistically and/or experimentally. Thus, path analysis may be

described, in part, as the process of hypothesizing a model of causal (structural) relations among measured variables—as often represented in a *path diagram*—and then formally examining observed data relations for degree of fit with the initial hypotheses’ expected relations.

Most modern treatments of path analysis trace the technique’s beginnings to the work of the biometrist Sewell Wright (e.g., Wright, 1918), who first applied path analysis to correlations among bone measurements. The method was barely noticed in the social sciences until Otis Duncan (1966) and others introduced the technique in sociology. Spurred by methodological, computational, and applied developments mainly during the 1970s and 1980s by such scholars as K. G. Jöreskog, J. Ward Keesling, David Wiley, Dag Sörbom, Peter Bentler, and Bengt Muthén, countless applications have appeared throughout the social and behavioral sciences. For example, shifting focus from correlations to covariances allowed for the development of a formal statistical test of the fit between observed data and the hypothesized model. Furthermore, the possibility of including latent (i.e., unobserved) factors into path models and the development of more general estimation techniques (e.g., MAXIMUM LIKELIHOOD, asymptotically DISTRIBUTION FREE) addressed major initial limitations of traditional correlational path analysis. Continuous refinements in specialized computer software throughout the 1980s, 1990s, and today (e.g., LISREL, EQS, AMOS, Mplus) have led to a simultaneous increase in technical sophistication and ease of use and thus have contributed to the proliferation of path analysis and more general SEM applications over the past three decades.

In this entry, typical steps in the path analysis process are introduced theoretically and via example. These steps include model conceptualization, parameter identification and estimation, effect decomposition, data-model fit assessment, and potential model modification and cross-validation. More detail about path analysis and SEM in general may be found in, for example, Bollen (1989), Kaplan (2000), Kline (1998), and Mueller (1996).

FOUNDATIONAL PRINCIPLES

As an example, an education researcher might have a theory about the relations among four constructs: Mathematics Self-Concept (MSC), Reading Self-Concept (RSC), Task Goal Orientation (TGO),

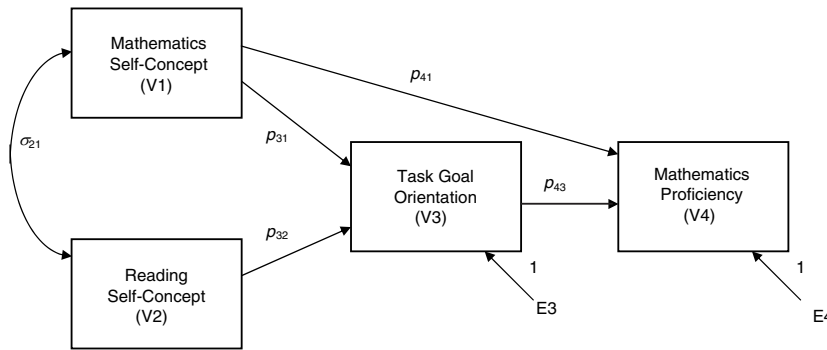


Figure 1 Hypothetical Path Model

and Mathematics Proficiency (MP). The first three constructs are operationalized as scores from rating scale instruments, whereas the latter is operationalized as a standardized test score. Using the measured variables as proxies for their respective constructs, the theory may be expressed as shown in Figure 1. Each box represents a measured variable, labeled as V. A single-headed arrow represents a formal theoretical statement that the variable at the tail of the arrow might have a direct causal bearing on the variable at the head of the arrow (but not vice versa). As seen in the figure, RSC is hypothesized to affect TGO directly, MSC to affect TGO and MP, and TGO to affect MP. Variables that have no causal inputs within a model are said to be *independent* or *exogenous* variables (MSC and RSC), whereas those with causal inputs are *dependent* or *endogenous* variables (TGO and MP). Two other single-headed arrows appear in the figure, one into TGO and the other into MP. These signify all unrelated residual factors (“Errors,” labeled E) affecting each endogenous variable. Finally, a two-headed arrow indicates that the variables could be related but for reasons other than one causing the other (e.g., both influenced by some other variable(s) external to this model). MSC and RSC are hypothesized to have such a relation. Note that the two-headed arrow does *not* symbolize reciprocal causation (when variables are hypothesized to have a direct or indirect causal bearing on each other). Such would be an example of a larger class of *nonrecursive models* that is beyond the scope of this entry (see Berry, 1984; Kline, 1998).

On each single-headed arrow is an unstandardized *path coefficient* for the population, p , indicating the direction and magnitude of the causal relation between the relevant variables. These are akin to, but not necessarily identical to, unstandardized partial regression

weights. With regard to the two-headed arrow, as it represents a covariance between two variables, the symbol appearing atop the arrow is the familiar σ . Using these symbols, we may express the relations from the figure in two ways: as *structural equations* and as *model-implied relations*. Structural equations are regression-type equations expressing each endogenous variable as a function of its direct causal inputs. Assuming variables are mean centered (thereby eliminat-

ing the need for intercept terms), the structural equations for the current model are as follows:

$$V3 = p_{31}V1 + p_{32}V2 + 1E3,$$

$$V4 = p_{41}V1 + p_{43}V3 + 1E4.$$

These equations, along with any noncausal relations contained in the model (i.e., two-headed arrows), have implications for the variances and covariances one should observe in the data. Specifically, if the hypothesized model is true in the population, then the algebra of EXPECTED VALUES applied to the structural equations yields the following model-implied variances (Var) and covariances (Cov) for the population:

$$\text{Var}(V1) = \sigma_1^2$$

$$\text{Var}(V2) = \sigma_2^2$$

$$\text{Var}(V3) = p_{31}^2\sigma_1^2 + p_{32}^2\sigma_2^2 + 2p_{31}\sigma_{21}p_{32} + \sigma_{E3}^2$$

$$\begin{aligned} \text{Var}(V4) = & p_{41}^2\sigma_1^2 + 2p_{43}p_{31}p_{41}\sigma_1^2 \\ & + 2p_{43}p_{32}\sigma_{21}p_{41} \\ & + p_{43}^2(p_{31}^2\sigma_1^2 + p_{32}^2\sigma_2^2 \\ & + 2p_{31}\sigma_{21}p_{32} + \sigma_{E3}^2) + \sigma_{E4}^2 \end{aligned}$$

$$\text{Cov}(V1, V2) = \sigma_{21}$$

$$\text{Cov}(V1, V3) = p_{31}\sigma_1^2 + p_{32}\sigma_{21}$$

$$\text{Cov}(V1, V4) = p_{41}\sigma_1^2 + p_{31}p_{43}\sigma_1^2 + p_{32}p_{43}\sigma_{21}$$

$$\text{Cov}(V2, V3) = p_{32}\sigma_2^2 + p_{31}\sigma_{21}$$

$$\text{Cov}(V2, V4) = p_{32}p_{43}\sigma_2^2 + p_{41}\sigma_{21}$$

$$\begin{aligned} \text{Cov}(V3, V4) = & p_{43}(p_{31}^2\sigma_1^2 + p_{32}^2\sigma_2^2 + 2p_{31}\sigma_{21}p_{32} \\ & + \sigma_{E3}^2) + p_{31}p_{41}\sigma_1^2 + p_{32}p_{41}\sigma_{21} \end{aligned}$$

These 10 equations constitute the covariance structure of the model, with 9 unknown population parameters appearing on the right side (p_{31} , p_{32} , p_{41} , p_{43} , σ_{21} , σ_1^2 , σ_2^2 , σ_{E3}^2 , σ_{E4}^2). If one had the population variances and covariances among the four variables, those numerical values could be placed on the left, and the unknowns on the right could be solved for (precisely, if the hypothesized model is correct). Given only sample (co)variances, such as those shown in Table 1 for a hypothetical sample of $n = 1,000$ ninth graders, they may similarly be placed on the left, and estimates of the parameters on the right may be derived. To this end, each parameter in a model must be expressible as a function of the (co)variances of the observed variables. When a system of such relations can be uniquely solved for the unknown parameters, the model is said to be *just-identified*. When multiple such expressions exist for one or more parameters, the model is *overidentified*; in this case, a best-fit (although not unique) estimate for each parameter is derived. If, however, at least one parameter cannot be expressed as a function of the observed variables' (co)variances, the model is *underidentified*, and some or all parameters cannot be estimated on the basis of the data alone. This underidentification might be the result of the researcher attempting to impose a model that is too complex (i.e., too many parameters to be estimated) relative to the number of (co)variances of the observed variables. Fortunately, underidentification is rare in most path analysis applications, occurring only with less common model features (e.g., nonrecursive relations).

Table 1 Sample Matrix of Variances and Covariances

	V1	V2	V3	V4
V1 (mathematics self-concept)	1.951			
V2 (reading self-concept)	-.308	1.623		
V3 (task goal orientation)	.262	.242	1.781	
V4 (mathematics proficiency)	28.795	-4.317	17.091	1375.460

The model in Figure 1 is overidentified with 10 observed (co)variances and 9 parameters to be estimated: 1 covariance, 4 direct paths, 2 variable variances, and 2 error variances. The model thus has $10 - 9 = 1$ degree of freedom (*df*). Given that a

model is just- or (preferably) overidentified, parameter estimates can be obtained through a variety of estimation methods. These include maximum likelihood and generalized least squares, both of which assume multivariate normality and are asymptotically equivalent as well as asymptotically distribution-free estimation methods that generally require a substantially larger sample size. These methods iteratively minimize a function of the discrepancy between the observed (co)variances and those reproduced by substitution of iteratively changing parameter estimates into the model-implied relations. Using any SEM software package, the maximum likelihood estimates for the key paths in the current model are found to be (in the variables' original units) as follows: $\hat{p}_{31} = .163$, $\hat{p}_{32} = .180$, $\hat{p}_{41} = 13.742$, $\hat{p}_{43} = 7.575$, and $\hat{\sigma}_{21} = -.308$ (all of which are statistically significant at a .05 level). Standardized path coefficients, similar to beta weights in multiple regression, would be .170, .172, .518, .273, and -.173, respectively. Before focusing on these main parameter estimates, however, we should consider whether there is any evidence suggesting data-model misfit and any statistical—and theoretically justifiable—rationale for modifying the hypothesized model.

DATA-MODEL FIT ASSESSMENT AND MODEL MODIFICATION

One of the advantages of modern path analysis is its ability to assess the quality of the fit between the data and the model. A multitude of measures exists that assist the researcher in deciding whether to reject or tentatively retain an a priori specified overidentified model. In general, measures to assess the fit between the (co)variances observed in the data and those reproduced by the model and its parameter estimates can be classified into three categories: absolute, parsimonious, and incremental. Absolute fit indices are those that improve as the discrepancy between observed and reproduced (co)variances decreases, that is, as the number of parameters approaches the number of nonredundant observed (co)variances. Examples of such measures include the model chi-square statistic that tests the stringent null hypothesis that there is no data-model misfit in the population, the standardized root mean square residual (SRMR) that roughly assesses the average standardized discrepancy between observed and reproduced (co)variances, and the goodness-of-fit index (GFI) designed to evaluate the

amount of observed (co)variance information that can be accounted for by the model.

Parsimonious fit indices take into account not just the overall absolute fit but also the degree of model complexity (i.e., number of parameters) required to achieve that fit. Indices such as the adjusted goodness-of-fit index (AGFI), the parsimonious goodness-of-fit index (PGFI), and the root mean square error of approximation (RMSEA) indicate greatest data-model fit when data have reasonable absolute fit and models are relatively simple (i.e., few parameters). Finally, incremental fit indices such as the normed fit index (NFI) and the comparative fit index (CFI) gauge the data-model fit of a hypothesized model relative to that of a baseline model with fewer parameters.

The three types of fit indices together help the researcher to converge on a decision regarding a path model's acceptability. Hu and Bentler (1999) suggested joint criteria for accepting a model, such as CFI values of 0.96 or greater together with SRMR values less than 0.09 (or with RMSEA values less than 0.06). In the current example, the nonsignificant model chi-square of 2.831 (with 1 *df*) indicates that the observed (co)variances could reasonably occur if our model correctly depicted the true population relations. Although bearing good news for the current model, this statistic is typically ignored as it is notoriously sensitive to very small and theoretically trivial model misspecifications under large sample conditions. As such, other fit indices are generally preferred for model evaluation. For the current model, no appreciable data-model inconsistency is suggested when applying the standards presented above: CFI = 0.997, SRMR = 0.013, and RMSEA = 0.043.

After the data-model fit has been assessed, a decision about that model's worth must be reached. Acceptable fit indices usually lead to the conclusion that no present evidence exists warranting a rejection of the model or the theory underlying it. This is not to say that the model and theory have been confirmed, much less proven as correct; rather, the current path model remains as one of possibly many that satisfactorily explain the relations among the observed variables. On the other hand, when fit indices suggest a potential data-model misfit, one might be reluctant to dismiss the model entirely. Instead, attempts are often made to modify the model post hoc, through the addition of paths, so that acceptable fit indices can be obtained. Most path analysis software packages will facilitate such model "improvement" by providing

modification indices (Lagrange multiplier tests) indicating what changes in the model could reap the greatest increase in absolute fit, that is, decrease in the model chi-square statistic. Although such indices constitute a potentially useful tool for remedying incorrectly specified models, it seems imperative to warn against an atheoretical hunt for the model with the best fit. Many alternative models exist that can explain the observed data equally well; hence, attempted modifications must be based on a sound understanding of the specific theory underlying the model. Furthermore, when modifications and reanalyses of the data are based solely on data-model misfit information, subsequent fit results might be due largely to chance rather than true improvements to the model. Modified structures therefore should be cross-validated with an independent sample whenever possible. From the current analysis, none of the modification indices suggested changes to the model that would result in a statistically significant improvement in data-model fit (i.e., a significant decrease in the model chi-square statistic); indeed, as the overall model chi-square value is 2.831 (significance level > 0.05), there is no room for statistically significant improvement. Thus, no model modification information is reported here.

At last, given satisfactory data-model fit, one may draw conclusions regarding the specific model relations. Interpretation of path coefficients is similar to that of regression coefficients but generally with a causal bent. For example, $\hat{p}_{31} = 0.163$ implies that, under the hypothesized model, a 1-unit increase in scores on the MSC scale directly causes a 0.163-unit increase in scores on the TGO scale, on average, holding all else constant. Recall that this was a statistically significant impact. Alternatively, one may interpret the standardized path coefficients, such as the corresponding value of 0.170. This implies that, under the hypothesized model, a 1 standard deviation increase in scores on the MSC scale directly causes a 0.170 standard deviation increase in scores on the TGO scale, on average, holding all else constant. Other direct paths, unstandardized and standardized, are interpreted similarly as the *direct effect* of one variable on another under the hypothesized model.

In addition to the direct effects, other *effect coefficients* may be derived as well. Consider the modeled relation between RSC and MP. RSC does not have a direct effect on MP, but it does have a direct effect on TGO, which, in turn, has a direct effect on MP. Thus, RSC is said to have an *indirect effect* on MP

because, operationally speaking, one may follow a series of two or more single-headed arrows flowing in the same direction to get from one variable to the other. The magnitude of the indirect effect is the product of its constituent direct effects, $(0.180)(7.575) \approx 1.364$ for the unstandardized solution. This unstandardized indirect effect value implies that a 1-unit increase in scores on the RSC scale indirectly causes a 1.364-unit increase in scores on the MP test, on average, holding all else constant. Similarly, for the standardized solution, the standardized indirect effect is $(0.172)(0.273) \approx 0.047$, which implies that a 1 standard deviation increase in scores on the RSC scale indirectly causes a .047 standard deviation increase in scores on the MP test, on average, holding all else constant.

Finally, notice that MSC actually has both a direct effect and an indirect effect on MP. In addition to the unstandardized and standardized direct effects of 13.742 and 0.518, respectively, the indirect effect may be computed as $(0.163)(7.575) \approx 1.235$ in the unstandardized metric and $(0.170)(0.273) = 0.046$ in the standardized metric. This implies that the total causal impact of MSC on MP, the *total effect*, is the sum of the direct and indirect effects. For the unstandardized solution, this is $13.742 + 1.235 \approx 14.977$, implying that a 1-unit increase in scores on the MSC scale causes in total a 14.977-unit increase in scores on the MP test, on average, holding all else constant. Similarly, for the standardized solution, this is $0.518 + 0.046 = 0.564$, implying that a 1 standard deviation increase in scores on the MSC scale causes in total a 0.564 standard deviation increase in scores on the MP test, on average, holding all else constant. Thus, through this type of *effect decomposition*, which uses the full context of the hypothesized model, one can learn much more about the relations among the variables than would be provided by the typically atheoretical correlations or covariances alone.

SUMMARY

Path analysis has become established as an important analysis tool for many areas of the social and behavioral sciences. It belongs to the family of structural equation modeling techniques that allow for the investigation of causal relations among observed and latent variables in a priori specified, theory-derived models. The main advantage of path analysis lies in its ability to aid researchers in bridging the often-observed gap between theory and observation. As has

been highlighted in this entry, path analysis is best understood as a process starting with the theoretical and proceeding to the statistical. If the theoretical underpinnings are ill-conceived, then interpretability of all that follows statistically can become compromised. Researchers interested in more detail regarding path analysis, as well as its extension to the larger family of SEM methods, are referred to such resources as Bollen (1989), Kaplan (2000), Kline (1998), and Mueller (1996). Parallel methods involving categorical variables, which have foundations in LOG-LINEAR MODELING, may also be of interest (see Goodman, 1973; Hagenaars, 1993).

—Gregory R. Hancock and Ralph O. Mueller

REFERENCES

- Berry, W. D. (1984). *Nonrecursive causal models*. Beverly Hills, CA: Sage.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley.
- Duncan, O. D. (1966). Path analysis: Sociological examples. *American Journal of Sociology*, 72, 1–16.
- Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika*, 60, 179–192.
- Hagenaars, J. A. (1993). *Loglinear models with latent variables*. Newbury Park, CA: Sage.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1–55.
- Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.
- Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.
- Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS*. New York: Springer.
- Wright, S. (1918). On the nature of size factors. *Genetics*, 3, 367–374.

PATH COEFFICIENT. See PATH ANALYSIS

PATH DEPENDENCE

Originally formulated for use in economics, the concept of path dependence now appears with increasing

frequency in other social sciences, notably historical sociology and political science. It is used to explain how certain outcomes are the result of a particular sequence of events and how that unique sequence constrains future options.

When an outcome is the result of path dependency, some general factors about the outcome's history can be observed. First, there is an aspect of unpredictability. This means that despite all prior knowledge of a certain hoped-for outcome, some events will take place that could not have been foreseen—they will be historically contingent. Second, path dependence emphasizes the importance of small events. Economist W. Brian Arthur (1994) notes that “many outcomes are possible, and heterogeneities, small indivisibilities, or chance meetings become magnified by positive feedbacks to ‘tip’ the system into the actual outcome” (p. 27). Most important, however, is the third factor, which is “lock-in.” As small, seemingly inconsequential decisions are made along an outcome's path, these decisions are constantly being reinvented and reinforced. This is the logic of increasing returns. Once a decision is made (about a new technology or an institution, for example), that initial decision becomes very difficult to change. It is “locked in”—because it is constantly used and revised. In this way, path-dependent processes are unforgiving: After a decision is made, it is very costly to change and becomes even more costly as time passes. This persistence often leads to suboptimal outcomes, another hallmark of a path-dependent trajectory.

Illustrations of path dependence are useful to fully understand the concept. Economist Paul David (1985) first described path dependence in a study of the “QWERTY” keyboard arrangement. David argued that the now familiar arrangement of keys on a typewriter was the result of a random decision, and once typewriters were manufactured with that particular arrangement and people began to learn to use it, the arrangement became too costly to reverse. Although there were likely many alternatives (some of them perhaps more efficient) to the QWERTY keyboard, the initial decision to use that arrangement became “locked in” over time, and more and more people used it. As a result, changing the arrangement of keys now is virtually unthinkable.

A frequent criticism of explanations that employ path dependence is that they simply assert the importance of history. Paul Pierson (2000) emphatically notes, however, that “it is not the past *per se* but the unfolding of events over time that is theoretically central” (p. 264). This sentiment is echoed by other

proponents of path dependence, notably James Mahoney. Mahoney (2000) offers a framework for path-dependent analysis, suggesting that all path-dependent arguments must demonstrate (a) that the outcome under investigation was particularly sensitive to events occurring early in the historical sequence, (b) that these early events were historically contingent (unpredictable), and (c) that the historical events must render the “path” to the outcome relatively inertial or DETERMINISTIC (Mahoney, 2000, pp. 510–511). Such an analysis allows for a clear explanation of how, for example, certain institutions emerge, persist, and become resistant to change.

—Tracy Hoffmann Slagter

REFERENCES

- Arthur, W. B. (1994). *Increasing returns and path dependence in the economy*. Ann Arbor: University of Michigan Press.
- David, P. (1985). Clio and the economics of QWERTY. *American Economic Review*, 75, 332–337.
- Mahoney, J. (2000). Path dependence in historical sociology. *Theory and Society*, 29, 507–548.
- Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. *American Political Science Review*, 94, 251–267.

PATH DIAGRAM. See PATH ANALYSIS

PEARSON'S CORRELATION COEFFICIENT

The Pearson's correlation coefficient (denoted as r_{xy}) best represents the contemporary use of the SIMPLE CORRELATION that assesses the LINEAR relationship between two variables. The coefficient indicates the strength of the relationship, with values ranging from 0 to 1 in absolute value. The larger the magnitude of the coefficient, the stronger the relationship between the variables. The sign of the coefficient indicates the direction of the relationship as null, positive, or negative. A null relationship between variables X and Y suggests that an increase in variable X is accompanied with both an increase and a decrease in variable Y and vice versa. A positive relationship indicates that an increase (or decrease) in variable X is associated with an increase (or decrease) in variable Y . In contrast, a negative relationship suggests that an increase (or

decrease) in variable X is associated with a decrease (or increase) in variable Y .

ORIGIN OF PEARSON'S CORRELATION COEFFICIENT

In the 19th century, Darwin's theory of evolution stimulated interest in making various measurements of living organisms. One of the best-known contributors of that time, Sir Francis Galton, offered the statistical concept of correlation in addition to the law of heredity. In his 1888 article (cited in Stigler, 1986), "Co-Relations and Their Measurement, Chiefly from Anthropometric Data," Galton demonstrated that the REGRESSION lines for the lengths of the forearm and head had the same slope (denoted as r) when both variables used the same scale for the measurement. He labeled the r as an index of *co-relation* and characterized its identity as a REGRESSION COEFFICIENT. The spelling of *co-relation* was deliberately chosen to distinguish it from the word *correlation*, which was already being used in other disciplines, although the spelling was changed to *correlation* within a year. However, it was not until Karl Pearson (1857–1936) that Pearson's correlation coefficient was introduced.

COMPUTATIONAL FORMULAS

The Pearson's correlation coefficient is also called the Pearson product-moment correlation, defined by

$$\frac{\sum z_{X_i} z_{Y_i}}{n},$$

because it is obtained by multiplying the z -SCORES of two variables (i.e., the products), followed by taking the average (i.e., the moment) of these products. Various computational formulas for Pearson's correlation coefficient can be derived from the above formula (Chen & Popovich, 2002), as demonstrated as follows:

$$\begin{aligned} \frac{\sum z_{X_i} z_{Y_i}}{n} &= \frac{\sum \left(\frac{(X_i - \bar{X})}{s_X} \right) \left(\frac{(Y_i - \bar{Y})}{s_Y} \right)}{n} \\ &= \frac{\frac{1}{s_X s_Y} \sum (X_i - \bar{X})(Y_i - \bar{Y})}{n} \\ &= \frac{\sum \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n}}{s_X s_Y} \\ &= \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sqrt{\left(\sum X_i^2 - \frac{(\sum X_i)^2}{n} \right) \left(\sum Y_i^2 - \frac{(\sum Y_i)^2}{n} \right)}}. \end{aligned}$$

INTERPRETATION OF PEARSON'S CORRELATION COEFFICIENT

Pearson's correlation coefficient does not suggest a cause-and-effect relationship between two variables. It cannot be interpreted as a proportion, nor does it represent the proportionate strength of a relationship. For example, a correlation of .40 is not twice as strong a relationship compared to a correlation of .20. In contrast to Pearson's correlation coefficient, r_{xy}^2 (also labeled as the COEFFICIENT OF DETERMINATION) should be interpreted as a proportion. Say the Pearson's correlation coefficient for class grade and intelligence is .7. The r_{xy}^2 value of .49 suggests that 49% of the variance is shared by intelligence and class grade. It can also be interpreted that 49% of the variance in intelligence is explained or predicted by class grade and vice versa.

SPECIAL CASES OF PEARSON'S CORRELATION COEFFICIENT

There are several well-known simple correlations such as the point-biserial correlation coefficient, the phi coefficient, the SPEARMAN CORRELATION COEFFICIENT, and the ETA COEFFICIENT. All of them are special cases of Pearson's correlation coefficient. Suppose we use Pearson's correlation coefficient to assess the relationship between a dichotomous variable (e.g., gender) and a continuous variable (e.g., income). This correlation is often referred to as the point-biserial correlation coefficient. We can also use it to gauge the relationship between two dichotomous variables (e.g., gender and infection), and we refer to this coefficient as the Phi coefficient. If both variables are measured at the ORDINAL level and ranks of their characteristics are assigned, Pearson's correlation coefficient can also be applied to gauge the relationship. This coefficient is often labeled as the *Spearman correlation coefficient*. Finally, Pearson's correlation coefficient, when applied to measure the relationship between a multi-chotomous variable (e.g., races, schools, colors) and a continuous variable, is referred to as the Eta coefficient (Wherry, 1984).

USES OF PEARSON'S CORRELATION COEFFICIENT

The Pearson's correlation coefficient can be used in various forms for many purposes. It can be used for, although is not limited to, any of the following functions: (a) describing a relationship between two

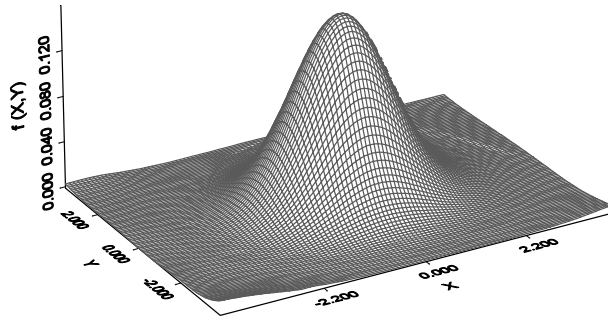


Figure 1 Bivariate Normal Distribution, Given $\rho = 0$ and $N = 1,000$

variables as a descriptive statistic; (b) examining a relationship between two variables in a POPULATION as an INFERENTIAL STATISTIC; (c) providing various RELIABILITY estimates such as CRONBACH'S ALPHA, TEST-RETEST RELIABILITY, and SPLIT-HALF RELIABILITY; (d) evaluating VALIDITY evidence; and (e) gauging the strength of effect for an intervention program. When Pearson's correlation coefficient is used to infer population RHO, the assumption of a bivariate NORMAL DISTRIBUTION is required. Both variables X and Y follow a bivariate normal distribution if and only if, for every possible linear combination W , where

$W = c_1X + c_2Y$, the distribution of W is normal, with neither c values being zero (Hays, 1994). An example of a bivariate normal distribution when $\rho = 0$ is depicted in Figure 1.

SAMPLING DISTRIBUTION OF PEARSON'S CORRELATION COEFFICIENT

The shape of the SAMPLING DISTRIBUTION for Pearson's correlation coefficient varies as a function of population rho (ρ) and sample size, although the impact of sample size is minimal when it is large. When population ρ equals zero, the sampling distribution is symmetrical and distributed as a T -DISTRIBUTION, as depicted in Figure 2. As seen below, the sampling distribution—based on 1,000 randomly generated samples, given $\rho = 0$ and $n = 20$ —is overall symmetrical. Therefore, to test the NULL HYPOTHESIS of $\rho = 0$, a T -TEST can be conducted, $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with DEGREES OF FREEDOM of $n - 2$.

However, the sampling distribution becomes negatively SKEWED when $|\rho|$ increases, given small sample sizes, as seen in Figure 3. The distribution can be made normal if each Pearson's correlation coefficient in the distribution is TRANSFORMED into a new

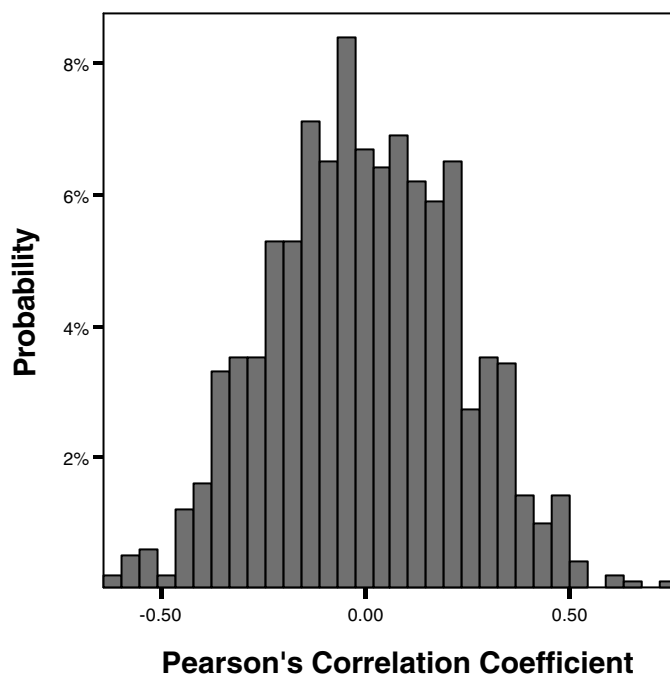


Figure 2 Sampling Distributions of 1,000 Pearson's Correlation Coefficients, Given $n = 20$ and $\rho = 0$

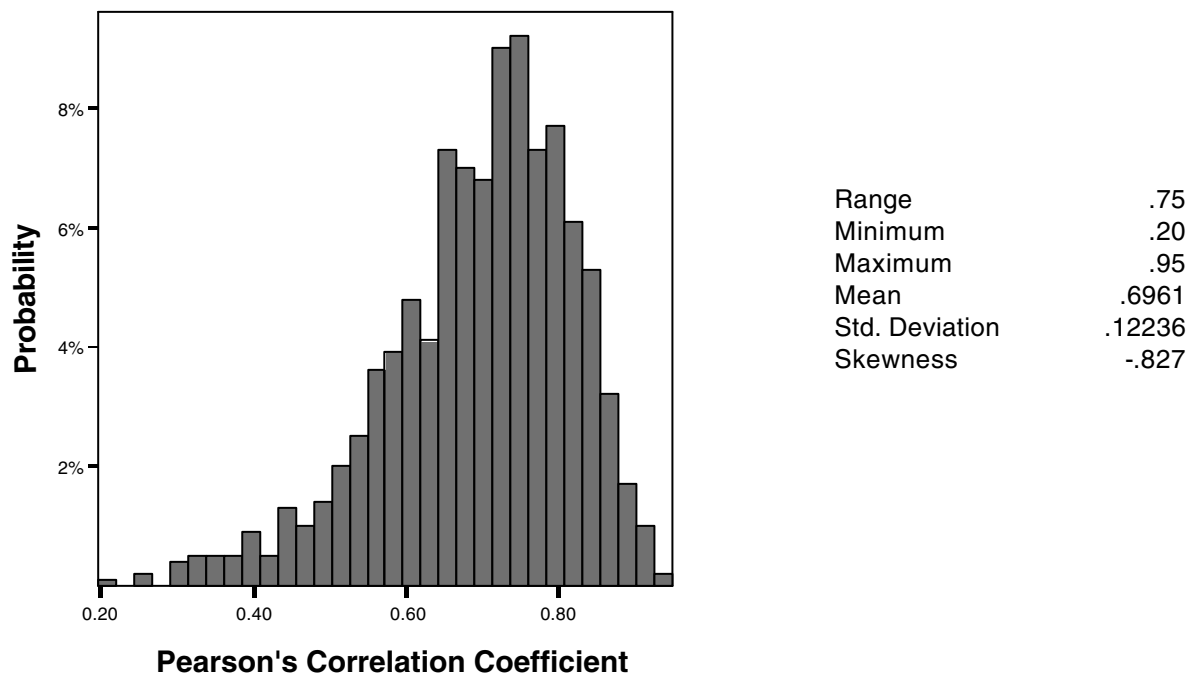


Figure 3 Sampling Distributions of 1,000 Pearson's Correlation Coefficients, Given $n = 20$ and $\rho = 0.7$

variable, labeled Fisher's z (denoted as z_r), based on $0.5 \times \log_e \left(\frac{1+r}{1-r} \right)$, where $\log_e$ is the natural LOGARITHM function. Consequently, to examine the null hypothesis that ρ equals any nonzero value (denoted as ρ_θ), we can conduct a z -test, $z = (z_r - z_\theta) \sqrt{(n-3)}$, where z_r is the transformed value of the Pearson's correlation coefficient for a sample, and z_θ is the transformed value of ρ_θ .

FACTORS THAT AFFECT PEARSON'S CORRELATION COEFFICIENT

The size of Pearson's correlation coefficient can be affected by many factors. First, the size of the coefficient can be either 1 or -1 only if the distributions of both variables are symmetrical and have the same form or shape. It can also be affected by the degree of linearity and small sample sizes, particularly when OUTLIERS exist. In addition, the size of the coefficient can either increase or decrease as a result of RANGE restriction. When the range of either variable is restricted on either end of the distribution, Pearson's correlation coefficient tends to decrease. In contrast, it tends to increase when the middle range of the distribution is restricted. Furthermore, the size of the coefficient may be misleading when it is calculated based on heterogeneous samples. Finally, it

can be influenced by measurement errors associated with the variables. The more ERRORS the measured variables contain, the smaller the correlation between the variables.

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage.
Hays, W. L. (1994). *Statistics* (5th ed.). New York: Harcourt Brace College.
Stigler, S. M. (1986). *The history of statistics*. Cambridge, MA: Belknap Press of Harvard University Press.
Wherry, R. J., Sr. (1984). *Contributions to correlational analysis*. New York: Academic Press.

PEARSON'S R. See PEARSON'S CORRELATION COEFFICIENT

PERCENTAGE FREQUENCY DISTRIBUTION

Percentage frequency distribution is a form of RELATIVE FREQUENCY DISTRIBUTION when the y-axis is

a percentage (instead of an absolute frequency) scale of cases subsumed by each category of a VARIABLE. For example, of the 22,000 undergraduate students registered at a certain American university during a certain spring semester, 30% were freshmen, 27% were sophomores, 20% were juniors, and 23% were seniors.

—Tim Futing Liao

See also RELATIVE FREQUENCY DISTRIBUTION

PERCENTILE

Percentiles may be defined as a system of measurement based on percentages as opposed to the absolute values of a variable. Percentages are based on a SCALE of 100 (i.e., there are 100 divisions), with each division being one percentage or percentile.



Figure 1 Division of a Line

The line in Figure 1 may be divided into 100 equal pieces. Halfway would be 50 equal divisions of the line. This position is 50% of the total length of the line and would also be known as the 50th percentile. If we stopped at the 75th division, this would be 75% of the total length of the line and would be the 75th percentile. If we stopped at the 62nd division, this would be 62% of the total length of the line and would be the 62nd percentile.

Clearly, there is no problem when the variable we are measuring is measured out of 100—the absolute measuring system and the percentage measuring system are one and the same thing. That is, the absolute measuring system has 100 equal divisions, as does the percentage measuring system. An exam score of 42 would represent a percentage score of 42% and be considered to be the 42nd percentile.

What happens when the two measuring systems are not the same? How would we work out percentiles if the maximum score on an examination paper were 30? The answer is to convert the absolute score into a percentage.

Example: Candidate 1 scores 25 marks out of 30 in a statistics examination.

We need to translate the actual 25 marks out of 30 into a percentage, and this is done using the following formulas:

$$\text{Percentage mark} = (\text{actual mark}/\text{maximum mark}) \times 100/1.$$

From our example is the following (remember the rule of mathematics—perform the calculations inside the brackets first):

$$\begin{aligned} \text{Percentage mark} &= (25/30) \times 100/1 \\ &= 83.33\% \text{ or the } 83.33 \text{ percentile.} \end{aligned}$$

The score of 25 would be considered to be the 83rd percentile (rounded down to the nearest whole number). Both the candidate and the examiner are now in a better position regarding interpretation and what conclusions to draw.

We can take percentile analysis slightly further by RANKING the percentiles. To do this, we need to know the frequency attributable to the scores of the variable we are interested in. A percentile rank is the percentage of scores in a DISTRIBUTION that is equal to, or greater than, a specified absolute score.

If the examiner of the statistics paper knows that there were 40 students out of 46 who achieved a mark of 25 or more, then we can calculate the percentile RANK by changing the absolutes into percentages:

$$(40/46) \times 100/1 = 86.95\%.$$

That is, 86.95% of all students achieved a mark of 25 or more, and the student is in the 87th percentile (rounded up to nearest whole number). The percentile RANK would be 87.

By dividing the DISTRIBUTION of a VARIABLE into percentiles, it is now possible to analyze each percentile as a group, examining the similarities and differences between these groups. More often than not, we group the variables into DECILES and QUARTILES to produce such an analysis. Alternatively, by calculating the RANK percentile, we can indicate the RANK of a person RELATIVE to others in the study.

Let us assume we have collected data on salary, sex, expenditure on goods and services, and so forth. By dividing the salary range into percentiles, we could analyze each percentile group to look for differences and similarities regarding the sex of a respondent and his or her salary. In addition, we could also examine the relationship between salary and the patterns of expenditure on goods and services. By using DECILES and

QUARTILES, we can reduce the number of groups and widen the salary range and are more likely to identify similarities and differences.

Ranking percentiles allows us to see the position we occupy out of 100, with the 100th percentile being the highest. In the example of the student who scored 25 on the statistics examination (which equated to the 83rd percentile), this informs us that this student could be ranked 83rd in this examination. We could produce percentiles ranks for all students, thereby allowing us to see how one student compares with another. Alternatively, we could take an individual student and calculate rank percentiles for all of their examinations, allowing us to see the range of the individual student's performance. This would allow us to identify individual strengths and weaknesses.

—Paul E. Pye

REFERENCES

- Collican, H. (1999). *Research methods and statistics in psychology*. London: Hodder & Staughton.
- Saunders, M. N. K., & Cooper, S. A. (1993). *Understanding business statistics: An active learning approach*. London: DP Publications.

PERIOD EFFECTS

Individual human subjects and sometimes other entities, such as cities or nations, can be regarded as existing in two time dimensions, subjective and objective. The subjective time dimension for an individual begins at birth, and parallels to birth can be found for other entities. The objective time dimension is measured with respect to events external to the individual or entity, dating from the beginning of the universe or, at a more modest level, the beginning of a research study. We call the subjective time dimension *age* and the objective time dimension *history*. Entities whose subjective time begins at the same objective time are said to belong to the same *cohort*.

In LONGITUDINAL RESEARCH, patterns of change in variables and patterns of relationships among variables are studied with respect to both age and history. The impact of age on variables or relationships among variables, ignoring historical time, is called a *developmental effect* or *age effect*. Examples of age effects include language acquisition and development of other

cognitive skills. The impact of living at a certain historical time on variables or relationships among variables, ignoring age, is called a *period effect* or *historical effect*. Examples of period effects include the use of computers, cell phones, and other devices, which depend on the level of technology in the historical period rather than on the age of the individual in that period, and attitudes and behaviors shaped by political and military events, such as expressing more patriotic attitudes or displaying the American flag more frequently in the United States after the bombing of Pearl Harbor or the destruction of the World Trade Center. The impact of being a certain age at a certain period (in other words, of being born in a certain year) is called a *cohort effect*. One of the principal problems in longitudinal research is separating the impacts of age, period, and cohort (Menard, 2002). The problem arises because if we use age, period, and cohort as predictors and assume a linear relationship of each with the dependent variable, the predictors are perfectly collinear, related by the equation, $\text{period} = \text{age} + \text{cohort}$, when cohort is measured as year of birth.

Different approaches to the separation of period effects from age and cohort effects are reviewed in Menard (2002). As noted by Hobcraft, Menken, and Preston (1982), period effects provide a proxy—often a weak proxy—for an explanation that is not really a matter of “what year is it” but rather a matter of what events occurred in that year, or before that year, at some significant period in the past. As Menard (2002) suggests, to the extent that it is possible, period effects should be replaced in longitudinal analysis by the historical events or trends for which they are proxies. At the aggregate level, it may also be possible to separate period effects from age or cohort effects by examining age-specific indicators of the variables of interest. This is commonly done in social indicators research, for example, by using trends in infant mortality or rates of labor force participation for individuals 16 to 64 years old as indicators of trends in national health and employment conditions.

—Scott Menard

REFERENCES

- Hobcraft, J., Menken, J., & Preston, S. (1982). Age, period, and cohort effects in demography: A review. *Population Index*, 42, 4–43.
- Menard, S. (2002). *Longitudinal research* (2nd ed.). Thousand Oaks, CA: Sage.

PERIODICITY

Periodic variation is a trademark of time that has long mystified ordinary people, inspired philosophers, and frustrated mathematicians. Through thousands of years, famines have alternated with prosperity, war with peace, ice ages with warming trends, among others, in a way that hinted at the existence of cycles. A cycle is a pattern that is “continually repeated with little variation in a periodic and fairly regular fashion” (Granger, 1989, p. 93). With the help of trigonometric functions, the cyclical component of a time series z_t may be expressed as follows:

$$z_t = A \sin(2\pi f t + \varphi) + u_t,$$

where A represents the amplitude, f the frequency, and φ the initial phase of the periodic variation, with u_t denoting WHITE-NOISE error. The amplitude indicates how widely the series swings during a cycle (the distance from peak to valley), whereas the frequency f measures the speed of the cyclical movement:

$$f = 1/P,$$

where P refers to the period of a cycle, the number of time units it takes for a cycle to be completed. Hence a 4-year cycle would operate at a frequency of 0.25 (cycles per unit time). The shortest cycle that can be detected with any given data would have a period of two time units. It would correspond to a frequency of 0.5, which is the fastest possible one. Any cycle that is done in fewer than two time units is not detectable. The longer the cycle—that is, the more time units it takes—the lower the frequency. A time series without a cycle would register a frequency of zero, which corresponds to an infinite period.

SPECTRAL ANALYSIS provides the major tool for identifying cyclical components in time series (Gottman, 1981; Jenkins & Watts, 1968). The typical plot of a spectral density function features spectral density along the vertical axis and frequency (ranging from 0 to 0.5) along the horizontal axis. A time series with a 4-year cycle (a year being the time unit of the series) would show a density peak at a frequency of $1/4$ or 0.25. The problem is that few time series of interest are generated in such perfectly cyclical fashion that their spectral density functions would be able to tell us so.

Aside from strictly seasonal phenomena, many apparent cycles do not repeat themselves with constant regularity. Economic recessions, for example, do not

occur every 7 years. Nor do they hurt the same each time. Although boom and bust still alternate, the period of this “cycle” is not fixed. Nor is its “amplitude.” That lack of constancy undermines the utility of DETERMINISTIC MODELS such as sine waves. As an alternative, Yule (1927/1971) proposed the use of PROBABILISTIC models for time series with cycles whose period and amplitude are irregular. His pioneering idea was to fashion a second-order autoregressive model to mimic the periodic, albeit irregular, fluctuations of sunspot observations over a span of 175 years. Yule’s AR(2) model had the following form:

$$z_t = b_1 z_{t-1} - b_2 z_{t-2} + u_t.$$

With only two PARAMETERS and LAGS going no further than two, this model succeeded in estimating the periodicity of the sunspot fluctuations (10.6 years). The model also provided estimates of the RANDOM disturbances that continually affected the amplitude and phase of those fluctuations. BOX-JENKINS MODELING has extended the use of probability models for identifying and estimating periodic components in time series, with a special set of models for seasonal fluctuations.

—Helmut Norpoth

REFERENCES

- Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. Cambridge, UK: Cambridge University Press.
- Granger, C. W. J. (1989). *Forecasting in business and economics*. Boston: Academic Press.
- Jenkins, G. M., & Watts, D. G. (1968). *Spectral analysis and its applications*. San Francisco: Holden-Day.
- Yule, G. U. (1971). On a method of investigating periodicities in disturbed series with special reference to Wolfer’s sunspot numbers. In A. Stuart & M. Kendall (Eds.), *Statistical papers of George Udny Yule* (pp. 389–420). New York: Hafner. (Original work published 1927)

PERMUTATION TEST

The permutation test is a NONPARAMETRIC test using the resampling method. The basic procedure of a permutation test is as follows:

1. Set the hypothesis.
2. Choose a test statistic and compute the test statistic T for the original samples (T_0).

3. Rearrange or permute the samples and recompute the test statistic for the rearranged samples $(T_1, T_2, \dots, T_x)$. Repeat this process until you obtain the distribution of the test static for all the possible permutations or sufficiently many permutations by random rearrangement.
4. Compare T_0 with the permutation distribution of T_1 to T_x and calculate the p value. Then, judge whether to accept or reject the hypothesis.

The permutation test is similar to the BOOTSTRAP: Both are computer intensive and limited to the data at hand. However, the methods of resampling are different: In the permutation test, you will permute the samples to construct the distribution, but in the bootstrap, you take a series of samples with replacement from the whole samples.

When you need to compare samples from two populations, the only requirement of the permutation test is that under the null hypothesis, the distribution from which the data in the treatment group are taken is the same as the distribution from which the untreated sample is taken.

The permutation test provides an exact statistic, not an approximation. The permutation test is as powerful as the parametric test if the sample size is large, and it is more powerful if the sample size is small. Thus, it is useful if it is difficult to obtain large samples such as sociometric data in a small group.

Let us look at hypothetical data on a new medicine to control blood pressure. The effect of the medicine is compared with a placebo. The data are collected by using 20 people whose usual blood pressures are around 165. These 20 people are evenly divided into two groups (each group consists of 10 people). The first 10 people get the placebo, and the last 10 people get the new medicine. The data for the placebo group are {132, 173, 162, 168, 140, 165, 190, 199, 173, 165}, and the data for the treatment group are {150, 124, 133, 162, 190, 132, 158, 144, 165, 152}. Is there any evidence that the new medicine is really effective in reducing blood pressure? Let us take the mean of both groups. For the placebo group, the mean is 166.7, and for the treatment group, it is 151.0. The difference of the means is 15.7. Now, we permute the data of 20 people. However, there are

$$\binom{20}{10} = 184,756$$

different combinations. We do not have to be concerned with all of those because dealing with this would be too time-consuming. So, we randomly permute the 20 data sufficiently many times—say, 100 times—and calculate the difference of the means of the first 10 and the last 10 permuted data. Then, we count the number of the cases such that the difference of the means of the original data, 15.7, is greater than the difference of the means of the permuted data. If we get 4 such cases out of 100 cases ($p = .04$), we conclude that the new medicine is effective in reducing blood pressure at $\alpha < .05$.

The permutation test can be applied in many social science research situations involving a variety of data; one such example is valued sociometric (network) data. Let us define the degree of symmetry between i and j as $s_{ij} = |A_{ij} - A_{ji}|$, where A is an adjacency matrix of a relation, and A_{ij} is a degree of the relation from i to j . The larger s implies that the relation deviates from symmetry or is less symmetric. Let us also define the symmetry in a whole group as

$$S = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j>i}^N s_{ij}.$$

We can calculate S , but we may wonder how large or small it is compared with symmetry in other groups because the distribution of S is unknown yet. So, we conduct a permutation test for symmetry with the test statistic S .

There are some practical guidelines for permutation tests dealing with sociometric data. First, we do not permute diagonal values because they usually hold a theoretically defined value such as 0 or 1. Second, we need to decide how to permute the matrix. For example, we may choose to permute the whole matrix, within each row, and so on, depending on your interest. In general, we can conduct a better test with a more restricted permutation. In the case of symmetry S , permutation within each row may make more sense because we calculate s (thus S), keeping each individual's properties such as the means and the variances.

For symmetry and transitivity tests, a computer program PermNet is available at the following Web site: www.meijigakuin.ac.jp/~rtsuji/en/software.html.

You can execute permutation tests with common statistics software packages. A macro is available for SAS (www2.sas.com/proceedings/sugi27/p251-27).

pdf). The Exact Test package of SPSS, StatXact, and LogXact will also do the test.

—Ryuhei Tsuji

REFERENCES

- Boyd, J. P., & Jonas, K. J. (2001). Are social equivalences ever regular? Permutation and exact tests. *Social Networks*, 23, 87–123.
- Good, P. (1994). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. New York: Springer.
- Tsuji, R. (1999). Trusting behavior in prisoner's dilemma: The effects of social networks. *Dissertation Abstracts International Section A: Humanities & Social Sciences*, 60(5-A), 1774. (UMI No. 9929113)

PERSONAL DOCUMENTS. See
DOCUMENTS, TYPES OF

PHENOMENOLOGY

In its most influential philosophical sense, the term *phenomenology* refers to descriptive study of how things appear to consciousness, often with the purpose of identifying the essential structures that characterize experience of the world. In the context of social science methodology, the term usually means an approach that pays close attention to how the people being studied experience the world. What is rejected here is any immediate move to evaluate that experience (e.g., as true or false by comparison with scientific knowledge) or even causally to explain why people experience the world the way that they do. From a phenomenological viewpoint, both these tendencies risk failing to grasp the complexity and inner logic of people's understanding of themselves and their world. Phenomenological sociology shares something in common with anthropological relativism, which insists on studying other cultures "in their own terms"; and with SYMBOLIC INTERACTIONIST approaches, which emphasize that people are constantly *making* sense of what happens, and that this process generates diverse perspectives or "worlds."

The phenomenological movement in philosophy was founded by Edmund Husserl at the end of the

19th century and was later developed in diverse directions by, among others, Heidegger, Merleau-Ponty, Sartre, and (most influentially of all for social science) Schutz. The fundamental thesis of phenomenology is the intentionality of consciousness: that we are always conscious *of* something. The implication is that the experiencing subject and the experienced object are intrinsically related; they cannot exist apart from one another. The influence of phenomenology has been especially appealing to those social scientists who oppose both POSITIVISM's treatment of natural science as the only model for rational inquiry *and* all forms of speculative social theorizing. It serves as an argument against these, as well as an alternative model for rigorous investigation.

One of the phenomenological concepts that has been most influential in social science is what is referred to as the "natural attitude." Husserl argued that rather than people experiencing the world in the way that traditional EMPIRICISTS claimed, as isolated sense data, most of the time they experience a familiar world containing all manner of recognizable objects, including other people, whose characteristics and behavior are taken as known (until further notice). Although there are areas of uncertainty in our experience of the world, and although sometimes our expectations are not met, it is argued that these problems always emerge out of a taken-for-granted background. The task of phenomenological inquiry, particularly as conceived by later phenomenologists such as Schutz (1973), is to describe the constitutive features of the natural attitude: how it is that we come to experience the world as a taken-for-granted and shared reality. They argue that this must be explicated if we are to provide a solid foundation for science of any kind.

Phenomenology has probably had its most direct influence on social research through the work of Berger and Luckmann (1967)—which stimulated some kinds of social CONSTRUCTIONISM—and through ETHNOMETHODOLOGY.

—Martyn Hammersley

REFERENCES

- Berger, P., & Luckmann, T. (1967). *The social construction of reality*. Harmondsworth, UK: Penguin.
- Embree, L., Behnke, E. A., Carr, D., Evans, J. C., Huertas-Jourda, J., Kockelmans, J. J., et al. (Eds.). (1996). *The encyclopedia of phenomenology*. Dordrecht, The Netherlands: Kluwer.

- Hammond, M., Howarth, J., & Keat, R. (1991). *Understanding phenomenology*. Oxford, UK: Blackwell.
- Schutz, A. (1973). *Collected papers* (4 vols.). Dordrecht, The Netherlands: Kluwer.

PHI COEFFICIENT. *See* SYMMETRIC MEASURES

PHILOSOPHY OF SOCIAL RESEARCH

The philosophy of social research is a branch of the PHILOSOPHY OF SOCIAL SCIENCE. It involves a critical examination of social research practice, as well as the various components of social research and their relationships.

The social researcher is confronted with a plethora of choices in the design and conduct of social research. These include the research problem; the RESEARCH QUESTION(S) to investigate this problem; the research strategies to answer these questions; the approaches to social enquiry that accompany these strategies; the concepts and theories that direct the investigation; the sources, forms, and types of data; the methods for making selections from these data sources; and the methods for collecting and analyzing the data to answer the research questions (Blaikie, 2000). At each point in the research process, philosophical issues are encountered in the evaluation of the options available and in the justification of the choices made. It is becoming increasingly difficult for social researchers to avoid these issues by simply following uncritically the practices associated with a particular PARADIGM.

PHILOSOPHICAL ISSUES

The philosophical issues that enter into these choices include types of ontological and epistemological assumptions, the nature of EXPLANATION, the types of THEORY and methods of construction, the role of language, the role of the researcher, OBJECTIVITY and truth, and generalizing results across time and space (Blaikie, 2000). As these issues are interrelated, the choice made on some issues has implications for the choice on other issues. This brief discussion of some of these issues is dependent on the more detailed discussion of related key concepts in this encyclopedia.

Ontological and Epistemological Assumptions

A key concern of philosophers of science and social science is the examination of claims that are made by theorists and researchers about the kinds of things that exist in the world and how we can gain knowledge of them. Social researchers, whether they realize it or not, have to make assumptions about both of these to proceed with their work. A range of positions can be adopted. To defend the position taken, social researchers need to have an understanding of the arguments for and against each position and be able to justify the stance taken. (See ONTOLOGY and EPISTEMOLOGY for a discussion of the range of positions.)

Theory Types and Their Construction

Philosophers have given considerable attention to the forms that theory takes and the processes by which theory is generated. In the social sciences, it is useful to distinguish between two types of theory: theoreticians' theory and researchers' theory. The former uses abstract concepts and ideas to understand social life at both the macro and micro levels. The latter is more limited in scope and is developed and/or tested in the context of social research. Although these two types of theory are related, their form and level of abstraction are different (Blaikie, 2000, pp. 142–143, 159–163). It is researchers' theory that is of concern here.

Some researchers' theories consist of general laws that are related in networks and are generated by INDUCTION from DATA. Others take the form of a set of propositions, of differing levels of generality, which form a deductive argument. Their origin is regarded as being unimportant, but their empirical testing is seen to be vital (see FALSIFICATIONISM). Some theories consist of descriptions of generative structures or mechanisms that are hypothesized to account for some phenomenon and are arrived at by the application of an informed imagination to a problem (see RETRODUCTION). Other theories consist of either rich descriptions or abstractions, perhaps in the form of ideal types of actors, courses of action, and situations. They are derived from social actors' concepts and interpretations (see ABDUCTION and INTERPRETIVISM). Philosophers of social science have put forward arguments for and criticisms of all four types of theory construction (see Blaikie, 1993).

The Nature of Explanation

One of the most vexed questions in the social sciences is how to explain social life. One controversy centers on the source of the explanation, whether it lies in social structures or in individual motives, that is, holism versus individualism. The issue is whether the source is to be found external to social actors or within them.

A related controversy is concerned with types of explanations, with the distinction between causal explanations and reason explanation. The former, which can take various forms, follows the logics of explanation used in the natural sciences, whereas the latter rejects these as being appropriate in the social sciences and replaces them with the reasons or motives social actors can give for their actions. Whether the use of reasons or motives can be regarded as explanations or just provide understanding is a matter of some debate.

Causal explanations, whether in the natural or social sciences, are of three major forms: pattern, deductive, and retroductive. In pattern explanations (see INDUCTION and POSITIVISM), a well-established generalization about regularities in social life can provide low-level explanations. For example, if it has been established that “juvenile delinquents come from broken homes,” it is possible to use this to explain why some young people commit crimes—it is their family background. Deductive explanations (see FALSIFICATIONISM) start with one or more abstract theoretical premises and, by the use of deductive logic and other propositions, arrive at a conclusion that can either be treated as a hypothesis or as the answer to a “why” research question. Retroductive explanations (see RETRODUCTION and CRITICAL REALISM) rely on underlying causal structures or mechanisms to explain patterns in observed phenomena.

The Role of Language

The philosophical issue here is what relationship there should be, if any, between scientific language and objects in the world. The positions adopted on this issue are dependent on the ontological assumptions adopted. For example, in POSITIVISM, the use of a theory-neutral observation language is considered to be unproblematic, whereas in FALSIFICATIONISM, this is rejected on the grounds that all observations are theory dependent.

In positivism, language is regarded as a medium for describing the world, and it is assumed that

there is an isomorphy between the structural form of language and the object-worlds to which language gives access. However, in some branches of INTERPRETIVISM, language is regarded as the medium of everyday, practical social activity. The issue here is the relationship between this lay language and social scientific language, with a number of writers advocating that the latter should be derived from the former by a process of ABDUCTION. The choice is between the imposition of technical language on the social world and the derivation of technical language from the everyday language.

The Role of the Researcher

Social researchers have adopted a range of stances toward the research process and the participants. These positions range from complete detachment to committed involvement. Philosophers of social science have critically examined these stances in terms of their achievability and their consequences for research outcomes.

Each stance is associated with the use of particular approaches to social enquiry and research strategies. The role of *detached observer* is concerned with OBJECTIVITY and is associated with positivism and the logics of induction and deduction. The role of *faithful reporter* is concerned with social actors' points of view and is associated with some branches of interpretivism. The role of *mediator of languages* is concerned with generating understanding and is associated with some branches of HERMENEUTICS and with the logic of abduction. The role of *reflective partner* is concerned with emancipation and is associated with both critical theory and feminism. Finally, the role of *dialogic facilitator* is associated with POSTMODERNISM and is concerned with reducing the researcher's authorial influence by allowing a variety of “voices” to be expressed (see Blaikie, 2000, pp. 52–56).

Objectivity

No concept associated with social research evokes both more emotion and confusion than objectivity. Researchers frequently aim to produce objective results and are often criticized when they are seen to fail. Some research methods are praised because they are seen to be objective, and others are criticized because they are assumed not to be objective. The philosophical issues are as follows: What does it mean to be “objective,”

and what do methods and results have to be like to be objective?

In positivism, being objective means keeping facts and values separate, that is, not allowing the researcher's values and prejudices to contaminate the data. This is a deceptively simple idea but one that turns out to be impossible to achieve. For a start, the notion of a "fact" is a complex one, and the difference between facts and values is not clear-cut. For some philosophers, keeping facts and values separate can mean not taking sides on issues involved in the research. However, some researchers favor a particular point of view (in feminism, women's points of view). Values themselves can also be a topic for investigation.

For some philosophers, objectivity is achieved by using methods that are regarded as being appropriate by a community of scientists. This, of course, means adopting a particular set of values. Objectivity is also considered by some to entail the use of replicability. Procedures are used that other researchers can repeat, thus making possible the validation of results. Again, acceptability of particular methods is the criterion.

Some interpretive social scientists eschew concerns with objectivity and argue that researchers should be as subjective as possible. What this means is that researchers should immerse themselves completely in the life of the people they study, as this is seen to be the only way another form of life can be understood. What they aim for is an authentic account of that form of life, an account that is faithful to social actors' points of view.

Truth

Philosophers have developed a number of theories to defend their views of the nature of "truth." The *correspondence* theory of truth appeals to evidence gained from "objective" observation. It is assumed that an unprejudiced observer will be able to see the world as it really is and will be able to report these observations in a theory-neutral language. It is assumed that a theory is true because it agrees with the facts (see POSITIVISM). Of course, to achieve this, there has to be agreement about what are the facts. The *tentative* theory of truth argues that the theory-dependence of observations and the nature of the logic of falsification mean that the testing of deductively derived theory can never arrive at the truth. Although finding the truth is the aim, the results of deductive theory testing are always subject to future revision. In contrast, the *consensus* theory

of truth is based on the idea of rational discussion free from all constraints and distorting influences. Although evidence may be important, it is rational argument that will determine the truth. Then there is the *pragmatic* theory of truth, which argues that something is true if it is useful, if it can be assimilated into and verified by the experiences of a community of people. In the correspondence theory, truth is seen to be established for all time. In the consensus theory, disinterested logical argument by anyone should arrive at the same truth. However, in the tentative theory, truth is seen to change as testing proceeds and, in the pragmatic theory, as experience changes. Finally, in the *relativist* theory of truth, there are many truths because it is assumed that there are multiple social realities. As no one has a privileged or neutral position from which to judge the various truths, it is impossible to determine which is the "truest." This range of positions makes the concept of truth rather less secure than common sense would suggest.

Generalizing

There seems to be a desire among practitioners of all sciences to be able to generalize their findings beyond the context in which they were produced, that their theories will apply across space and time (see GENERALIZATION). This is particularly true of positivism, in which inductive logic is used. Despite the fact that generalizations are based on finite observations, the aim is to produce universal laws, such as, "juvenile delinquents come from broken homes." Many philosophers have argued that this aim is impossible and have suggested less ambitious aspirations. Falsificationists, for example, accept the tentative nature of their theories, as well as the possibility that they may be replaced by better theories in the future. In addition, they may state some propositions in their theories in a general form but also include conditions that limit their applicability. In fact, it is the search for such conditions to which much scientific activity is devoted. For example, although the general proposition that "juvenile delinquents come from broken homes" may be included at the head of a deductive theory, limiting conditions—such as, "in some broken homes, children are inadequately socialized into the norms of society"—limit the theory to certain kinds of broken homes. The search for underlying mechanisms in CRITICAL REALISM is based on an assumption that they are real and not just the creation of scientists. However, such researchers see their task as

not only finding the appropriate mechanisms but also establishing the conditions and contexts under which they operate. Mechanisms may be general, but their ability to operate is always dependent on the circumstances. Hence, these theories are always limited in time and space. Within interpretivism, it is generally accepted that any social scientific account of social life must be restricted in terms of time and space. Any understanding must be limited to the context from which it was abductively derived. Whether the same type of understanding is relevant in other social contexts will be dependent on further investigation. In any case, all understanding is time bound.

One aspect of generalizing that causes considerable confusion is the difference between the form of theoretical propositions and what happens in sample surveys (see SURVEY). If a PROBABILITY SAMPLE is drawn from a defined POPULATION, and if response rates are satisfactory, the results obtained from the sample can be statistically generalized back to the population from which the sample was drawn. Beyond that population, all generalizations have to be based on other evidence and argument, not on statistical procedures. In this regard, generalizing beyond sampled populations is essentially the same as generalizing from NONPROBABILITY SAMPLES and CASE STUDIES (Blaikie, 2000).

ASPECTS OF SOCIAL RESEARCH

A number of aspects of social research practice are associated with the issues discussed in the previous section. The ones reviewed here involve choices from among alternatives.

Approaches to Social Enquiry

At an abstract level, social research is conducted within theoretical and methodological perspectives, that is, within different approaches to social enquiry. These approaches have been characterized in various ways and usually include positivism, falsificationism (or critical rationalism), hermeneutics, interpretivism, critical theory, critical realism, structuration theory, and feminism (Blaikie, 1993). Each approach works with its own combination of ontological and epistemological assumptions (see Blaikie, 1993, pp. 94–101). The choice of approach and its associated research strategy has a major bearing on other RESEARCH DESIGN

decisions and the conduct of a particular research project.

Research Strategies

Four basic strategies can be adopted in social research, based on the logics of induction, deduction, retroduction, and abduction (Blaikie, 1993, 2000). Each research strategy has a different starting point and arrives at a different outcome. The inductive strategy starts out with the collection of data, from which generalizations are made, and these can be used as elementary explanations. The deductive strategy starts out with a theory that provides a possible answer to a “why” research question. The conclusion that is logically deduced from the theory is the answer that is required. The theory is tested in the context of a research problem by the collection of relevant data. The retroductive strategy starts out with a hypothetical model of a mechanism that could explain the occurrence of the phenomenon under investigation. It is then necessary to conduct research to demonstrate the existence of such a mechanism. The abductive strategy starts out with lay concepts and meanings that are contained in social actors’ ACCOUNTS of activities related to the research problem. These accounts are then redescribed in technical language, in social scientific accounts. The latter provide an understanding of the phenomenon and can be used as the basis for more elaborate theorizing.

The four research strategies work with different ontological assumptions. The inductive and deductive strategies share the same ontological assumptions (see POSITIVISM, INDUCTION, FALSIFICATIONISM, and DEDUCTION); the retroductive strategy works with a tiered notion of reality, which particularly distinguishes between the domains of the actual (what we can observe) and the real (underlying structures and mechanisms) (see CRITICAL REALISM and RETRODUCTION). The abductive strategy adopts a social constructionist view of social reality (see INTERPRETIVISM and ABDUCTION). It is important for researchers to be aware of these differences as they have implications for other choices and for the interpretation of results.

The inductive strategy is only really useful for answering “what” research questions as generalizations from data have been shown to be an inadequate basis for answering “why” questions. The deductive and retroductive strategies are exclusively used for answering “why” questions, although both require that

descriptions of observed patterns first be established. The abductive strategy can be used to answer both “what” and “why” questions.

There are extensive philosophical criticisms of each strategy and debates about their relative merits. None of them is without some controversial features. In the end, the social researcher has to evaluate the arguments and then make a judgment about their suitability for the problem at hand while, at the same time, recognizing their limitations.

CONCLUSION

All this suggests that an understanding of the issues dealt with in the philosophy of social research is essential for any social scientist. The choices that have to be made in doing social research require social scientists to take sides on philosophical issues. A lack of understanding of what is involved in making these choices can lead to naive and inconsistent research practices.

—Norman Blaikie

REFERENCES

- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Blaikie, N. (2000). *Designing social research: The logic of anticipation*. Cambridge, UK: Polity.
- Hughes, J. A. (1990). *The philosophy of social research* (2nd ed.). Harlow, UK: Longman.
- Williams, M., & May, T. (1996). *Introduction to the philosophy of social research*. London: UCL Press.

PHILOSOPHY OF SOCIAL SCIENCE

The philosophy of social science is a branch of the philosophy of science, an important field of philosophy. The discipline of philosophy is difficult to define precisely. The original Greek meaning is the love of wisdom, but today it is associated with techniques of critical analysis. Although philosophy has no subject matter of its own and cannot produce new facts about the world, philosophers may be able to provide new understanding of known facts. They make the first principles of our knowledge clearer by exposing what is implicit in concepts and theories and exploring their implications. By analyzing the kind of language we use to talk about the world, philosophers can

help us to understand it. In particular, philosophers provide a critical reflection on our taken-for-granted interpretations of reality.

In the philosophy of science, critical analysis focuses on such topics as “what is reality” and “what constitutes knowledge of that reality,” on issues of ONTOLOGY and EPISTEMOLOGY. More specifically, this field of philosophy examines the nature of science, the logics of enquiry used to advance knowledge that is regarded as scientific, and the justifications offered for the views adopted.

The philosophy of social science is primarily concerned with three issues: the variety of ontological and epistemological assumptions that social researchers adopt in their theories and approaches to social enquiry, as well as the justifications given for these; the logics of enquiry that can be used to produce answers to RESEARCH QUESTIONS; and the appropriateness of the methods of data collection and analysis that are used within a particular logic of enquiry.

Many of the issues dealt with in the philosophy of science are also relevant in the philosophy of social science. However, the issue that particularly distinguishes the latter and that has divided philosophers and methodologists for about a hundred years is whether the methods or, more correctly, the logics of enquiry that have been developed and successfully used in the natural sciences should also be used in the social sciences. A problem with this debate is that there has also been a parallel debate about which logic of enquiry is the most appropriate in the natural sciences—INDUCTION, DEDUCTION, or some other procedure?

A fundamental choice for all social scientists is whether to use one of these logics of enquiry or one that is regarded by some philosophers as being more appropriate for the particular subject matter of the social sciences. The ontological assumptions adopted will determine whether social reality is seen to be essentially the same as natural reality or fundamentally different. In the latter case, a different logic (ABDUCTION) will have to be adopted, and this can restrict the choice of methods of data collection and analysis.

A central issue in this controversy is how to explain human action. Adopting the logics used in the natural sciences entails a concern with establishing causal explanations, with identifying the factors or mechanisms that produce those actions under particular conditions (see CAUSALITY). However, those social scientists who argue for a different kind of logic content

themselves with achieving some kind of understanding based on the reasons social actors give for their actions.

Hence, the choices that social scientists make about the kinds of research questions to investigate, as well as the logics of enquiry used to answer these questions, commit them to taking sides on profound and complex philosophical issues.

—Norman Blaikie

REFERENCES

- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Chalmers, A. (1982). *What is this thing called science?* St. Lucia, Australia: University of Queensland Press.
- O'Hear, A. (1989). *Introduction to the philosophy of science*. Oxford, UK: Clarendon.
- Rosenberg, A. (1988). *Philosophy of social science*. Oxford, UK: Clarendon.

PHOTOGRAPHS IN RESEARCH

Photographs have had a limited but important role in social science research. In recent years, this role has increased. Photography has had to overcome the fact that social science reasoning generally relies on the analysis of aggregate data, and photography, by definition, fixes attention on the particular. Second, the most vigorous focus on the visual aspects of society has been mustered by cultural studies and semiotics theorists, who have studied domination, resistance, and other cultural themes as in media, architecture, and other visual forms of society. This is not the use of photographs to gather data but rather analysis of the visually constructed world.

Photographs were first used in social science to catalogue aspects of material culture or social interaction. In the first 20 volumes of the *American Journal of Sociology*, between 1896 and 1916, sociologists used photographs to study such topics as the working conditions in offices and factories and living conditions in industrializing cities. As sociology sought a more scientific posture, the photographs were, for a half century, largely abandoned.

Anthropologists also used photography extensively in its formative decades to document and classify human groups. As anthropology abandoned paradigms

based on the taxonomic organization of racial classifications, photography declined in importance.

The full development of photography in ethnography is attributed to Gregory Bateson and Margaret Mead's (1942) study of Balinese culture. The authors studied social rituals, routine behavior, and material culture through extensive photographic documentation. Subsequently, photography has flourished in anthropological ETHNOGRAPHY.

Howard Becker was the first modern sociologist to link photography and sociology (Becker, 1974). Becker noted that photography and sociology, since their inception, had both been concerned with the critical examination of society. Early 20th-century photographers, such as Lewis Hine and Jacob Riis, documented urbanization and industrialization, the exploitation of labor, and the conditions of immigrants, all topics of concern to sociologists. Becker suggested that visual sociology should emerge from documentary practice engaged with sociological theory, and the sociological use of photographs developed in the context of concerns about VALIDITY, RELIABILITY, and SAMPLING.

In the meantime, sociologists demonstrated the utility of photographs to study social change (Rieger, 1996) and several other conventional sociological topics. Photographs, however, remain the exception rather than the rule in empirical studies. The challenge of using photographs to explore society at a macro or structural level is also beginning to be met in research such as Deborah Barndt's (1997) photo collage/field study of globalization and tomato production and consumption.

Researchers may employ photographs to elicit cultural information. This process, referred to as *photo elicitation*, uses photographs to stimulate culturally relevant reflections in interviews (Harper, 2002). The photographs may be made by the researcher during field observations, or they may come from historical or personal archives. In a recent study, for example (Harper, 2001), photos from an archive made 50 years ago that depicted the routines of work, community life, and family interaction were used to encourage elderly farmers to critically reflect on the meaning of change in an agricultural community.

Photo elicitation explores the taken-for-granted assumptions of both researcher and subject. Furthermore, because photo elicitation stimulates reflection based on the analysis of visual material, the

quality of memory and reflection is different than in words-alone interviews. Finally, photo elicitation challenges the authority of the researcher and redefines the research project to a collaborative search for meaning, rather than a mining of information. This places photo elicitation squarely in the emerging practices of “new ethnography” or “postmodern research methods.”

There are at least three ethical challenges in photographic social science. The first concerns the matter of anonymity, the second concerns the matter of INFORMED CONSENT, and the third represents the burden of “telling the visual truth.”

In conventional research, individuals are represented by words or, more commonly, as unidentified data in tables and graphs. Thus, the identity of research subjects is easily hidden. In the case of photographic research, however, it is often impossible to protect the PRIVACY of subjects. Thus, photographic research challenges a fundamental principle of social science research. The full resolution of this issue remains.

VISUAL RESEARCHERS face this problem partly through informed consent. When possible, subjects must be made aware of implications of publishing their likeness in social science research, and agreement to participate should be made clearly and in writing. Informed consent, however, is not possible when social scientific photographers make images in spaces such as public gatherings in parks, urban landscapes, and other sites of informal interaction. The full resolution of this problem also remains.

Visual researchers must maintain a firm commitment to “tell the visual truth” by not altering images to change their meaning and by presenting a visual portrait of reality that is consistent with the scientifically understood reality of the situation. This guiding principle is maintained as an overriding ethical consideration that overshadows the technical considerations listed above.

New methodologies, such as computer-based multimedia, the World Wide Web, CD publication, and virtual reality all offer possibilities for the innovative use of photography as a social science method. The future of photographs in social research will depend in part on how well these new technologies are integrated into social science methods and thinking. However, new journals devoted to visual studies and a growing list of books that use photographs as a primary form of data and communication

suggest that the day of the image is dawning in social science.

—Douglas Harper

REFERENCES

- Barndt, D. (1997). Zooming out/zooming in: Visualizing globalization. *Visual Sociology*, 12(2), 5–32.
- Bateson, G., & Mead, M. (1942). *Balinese character: A photographic analysis*. New York: New York Graphic Society.
- Becker, H. S. (1974). Photography and sociology. *Studies in the Anthropology of Visual Communication*, 1(1), 3–26.
- Harper, D. (2001). *Changing works: Visions of a lost agriculture*. Chicago: University of Chicago Press.
- Harper, D. (2002). Talking about pictures: A case for photo elicitation. *Visual Studies*, 17(1), 13–26.
- Rieger, J. (1996). Photographing social change. *Visual Sociology*, 11(1), 5–49.

PIE CHART

A pie chart is a circle segmented to illustrate shares or percentages of the total; each category appears as a “slice” of the “pie.” Bar charts can be used for the same purpose, but a pie chart gives a more readily comprehensible picture of how cases are “shared out” between categories, provided there are relatively few categories. (A HISTOGRAM or bar chart would be used in preference to a pie chart for showing trends across a variable—e.g., over time or by geographical locality.)

Figure 1, for example, illustrates what people living in a town center had to say about their area and what was good about it.

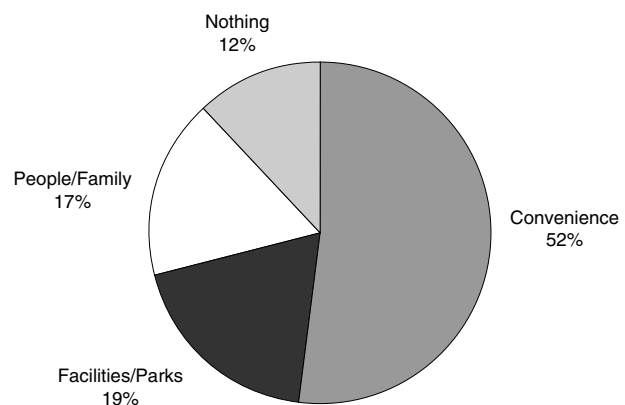


Figure 1 What Is Good About Living in Your Area?
SOURCE: Data from Hobson-West and Sapsford (2001).

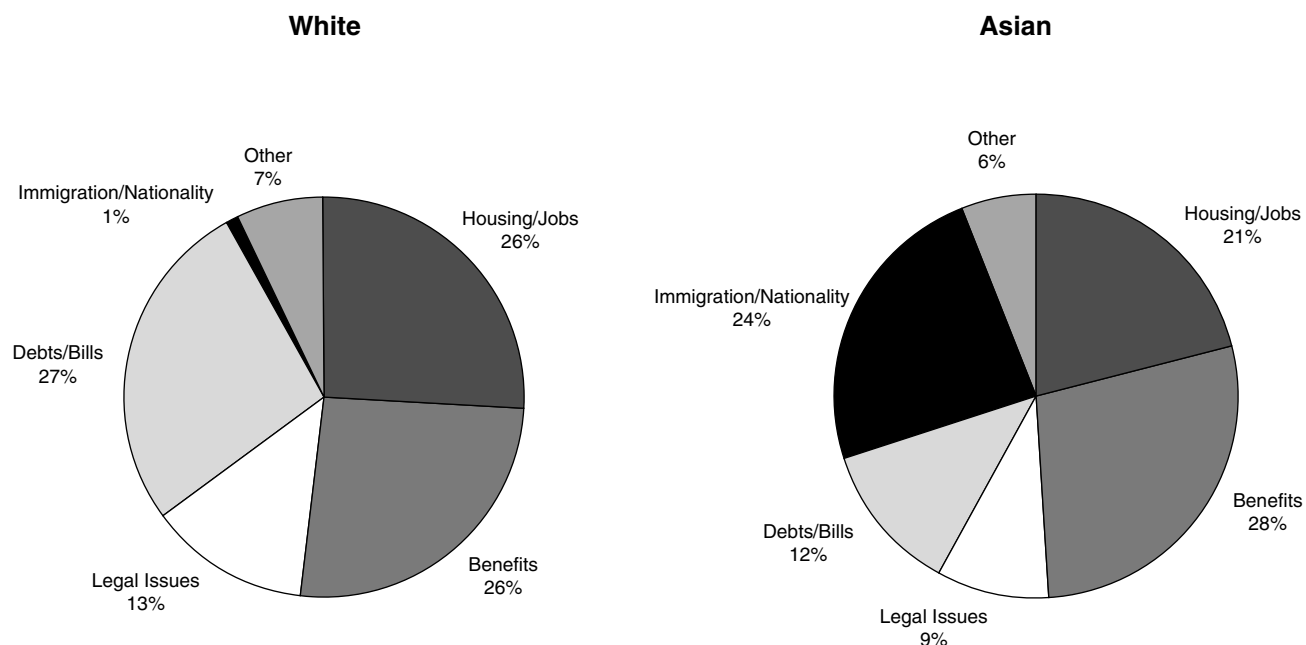


Figure 2 Purpose of Visit to the Citizens' Advice Bureau in Middlesbrough
 SOURCE: McGuinness and Sapsford (2002).

Pie charts can also be used to make simple comparisons between categories. Again, bar charts can be used for the same purpose and may, indeed, be preferable when there are several different groups to be compared, but pie charts give a graphic contrast between, say, two groups that can be seen easily and immediately. The “pies” in Figure 2, for instance, illustrate the difference in reasons for using a Citizens' Advice Bureau by ethnic group. The Citizens' Advice Bureaus are a network of volunteer offices across Britain that offer advice and guidance on dealing with the bureaucracy in the face of which the poor are otherwise powerless—with central and local government, taxation, landlords, shops, and other financial and legal matters—and they also offer counseling on debt and financial management. Figure 2 shows, unsurprisingly, that immigration and nationality are issues that are often brought up by the Asian population of Middlesbrough and very seldom by the White population. The White population is a little more likely than the Asian population to raise issues about housing and benefits (government support payments) and substantially more likely to have debt problems that they want to discuss.

—Roger Sapsford

REFERENCES

- Hobson-West, P., & Sapsford, R. (2001). *The Middlesbrough Town Centre Study: Final report*. Middlesbrough, UK: University of Teesside, School of Social Sciences.
- McGuinness, M., & Sapsford, R. (2002). *Middlesbrough Citizens' Advice Bureau: Report for 2001*. Middlesbrough, UK: University of Teesside, School of Social Sciences.

PIECEWISE REGRESSION.

See SPLINE REGRESSION

PILOT STUDY

Pilot study refers to

1. feasibility or small-scale versions of studies conducted in preparation for the main study and
2. the pretesting of a particular research instrument.

Good pilot studies increase the likelihood of success in the main study. Pilot studies should warn of possible project failures, deviations from PROTOCOLS, or problems with proposed methods or instruments and, it is hoped, uncover local politics or problems that may affect the research (see Table 1).

Table 1 Reasons for Conducting Pilot Studies

- Developing and testing adequacy of research instruments
- Assessing feasibility of (full-scale) study
- Designing research protocol
- Assessing whether research protocol is realistic and workable
- Establishing whether sampling frame and technique are effective
- Assessing likely success of proposed recruitment approaches
- Identifying possible logistical problems in using proposed methods
- Estimating variability in outcomes to determine sample size
- Collecting preliminary data
- Determining resources needed for study
- Assessing proposed data analysis techniques to uncover potential problems
- Developing research question and/or research plan
- Training researcher(s) in elements of the research process
- Convincing funding bodies that research team is competent and knowledgeable
- Convincing funding bodies that the study is feasible and worth funding
- Convincing stakeholders that the study is worth supporting

Pilot studies can be based on quantitative (QUANTITATIVE RESEARCH) and/or qualitative methods (QUALITATIVE RESEARCH), whereas large-scale studies may employ several pilots. The pilot of a QUESTIONNAIRE survey could begin with interviews or focus groups to establish the issues to be addressed in the questionnaire. Next, the wording, order of the questions, and the range of possible answers may be piloted. Finally, the research process could be tested, such as different ways of distributing and collecting the questionnaires, and precautionary procedures or safety nets devised to overcome problems such as poor recording and response rates (RESPONSE BIAS).

Difficulties can arise from conducting a pilot:

- Contamination of sample: Participants (RESPONDENTS) who have already been exposed to the pilot may respond differently

from those who have not previously experienced it.

- Investment in the pilot may lead researchers to make considerable changes to the main study, rather than decide that it is not possible.
- Funding bodies may be reluctant to fund the main study if the pilot has been substantial as the research may be seen as no longer original.

Full reports of pilot studies are relatively rare and often only justify the research methods and/or research tool used. However, the processes and outcomes from pilot studies might be very useful to others embarking on projects using similar methods or instruments. Researchers have an ethical obligation to make the best use of their research experience by reporting issues arising from all parts of a study, including the pilot phase.

—Edwin R. van Teijlingen and Vanora Hundley

REFERENCES

- De Vaus, D. A. (1995). *Surveys in social research* (4th ed.). Sydney, Australia: Allen & Unwin.
- Peat, J., Mellis, C., Williams, K., & Xuan, W. (2002). *Health science research: A handbook of quantitative methods*. London: Sage.
- van Teijlingen, E., & Hundley, V. (2002). The importance of pilot studies. *Nursing Standard*, 16(40), 33–36.

PLACEBO

This term *placebo* is used in connection with EXPERIMENTS. Its main use is in connection with medical investigations in which a CONTROL GROUP is led to believe that it is being given a treatment, such as a medicine, but the treatment is in fact a dummy one. The term is also sometimes used to refer to a treatment given to a control group that is not designed to have an effect. As such, its main role is to act as a point of comparison to the treatment or experimental group. A placebo effect occurs when people's behavior or state of health changes as a result of believing the placebo was meant to have a certain kind of effect, such as making them feel calmer.

—Alan Bryman

PLANNED COMPARISONS. See MULTIPLE COMPARISONS

POETICS

Poetics takes language both as its substance and as a conspicuous vehicle for wider ends of communication. It is a universal aspect of language-centered behavior that can be emphasized or de-emphasized according to the purposes at hand. The focus on composition and the conspicuous display of the language used in poetics are the keys to understanding it, not simply its forms of production (e.g., poetic prose, rich with metaphor and allegory; poetry as prose or verse; chanting, singing) or the kinds of messages sent (e.g., aesthetic, didactic, mythic, ritualized). It appears with increasing frequency today as a matter of special interest in the social sciences, especially in their concerns with texts as cultural artifacts. Rooted in part in the conflicts of POSTMODERNISM and in part in increasing awareness of the common denominators of cultural diversity, this “literary turn” has led to expanded research on semiotic behavior and the production of meaning in discourse, text construction and authority in performance, and, more generally, the place of poetics in the philosophical and critical problems associated with the representation and communication of experience in any form. It includes among its broad topics of inquiry the mutually constructed viewpoints of critical dialogics, the histories and forms of poetries in the world (from tribal societies to modern theater), studies of ritual and worldview and their relationships to language and culture, and the self-conscious filings of autoethnography and ethnographic fieldwork. It is rich with implications for social science.

ANTHROPOLOGICAL POETICS

The specific challenge of anthropological poetics includes developing sensitive and realistic cross-cultural translation skills and a genre of reporting that systematically incorporates edifying poetic qualities without sacrificing the essence of ethnographic accountability. We need to know something of the

substance and what it means to be situated differently in language and culture from the actor’s point of view while, at the same time, finding ways to combine and communicate those differences. On the assumption that all human beings are sufficiently alike to be open to discovery and mutually empathetic communications, anthropological poetics attempts to evoke a comparable set of experiences through the reader’s (or hearer’s) exposure to the text. Ethnographers have been attracted to these kinds of arguments precisely because they help to define the constraints on self-conscious research and writing. This increasing “literariness” has also led some social scientists to a classic obsession with lines, punctuation, and white spaces—in a word, to poetry as an alternative form of representation.

Poetry is by far the most celebrated category of poetic communication. Denying mind-body separations in concept and argument, it is cut very close to experiential ground. It takes place simultaneously in the deepest level of minding and in that part of consciousness where ideas and opinions are formed, constructing the poet as a cultural and sentient being even as the experience is unfolding. It is a self-revealing, self-constructing form of discovery, like writing in general and that fundamental human activity, “storying.” It takes its motivations and saturations from the exotic and the quotidian and from our dreams, creating and occupying sensuous and meaningful space in the process. Everything we do has meaning, from the simple act of digging up a sweet potato to the launching of rockets to the moon, and every meaning is ultimately anchored in the senses. Nevertheless, it is mostly only the poets who write about experience consistently from a sensual perspective—centering, decoding, reframing, and discoursing literally as “embodied” participant observers, full of touch, smell, taste, hearing, and vision, open to the buzz and joy, the sweat and tears, the erotics and anxieties of daily life. Poets offer their take on the world sometimes with a view closer than a bug’s eye on a plant, sometimes deep into the sublime, the universal. At work in ethnography, poets hope to reveal that world for what it is to themselves and to Others through mutual authorships, getting especially at the work that cannot take place within a single language or culture, as it is experienced in distinctive patterns and puzzles and can be shared with coparticipants, and yet doing so ultimately in a manner that makes the Otherness of obvious strangers collapse on close inspection. Poetry insists on being about all of us. But adding it to

the repertoire of social science methods is problematic. Among other things, it opens those disciplines up to the whole complicated realm of aesthetics and to what appears to the received wisdom to be a truckload of counterintuitive arguments about ethnographic representation, including the place of language (and writer consciousness) in scientific writing.

SCIENCE AND POETICS

Scientific inquiry can give us a more or less dispassionate glimpse of causal relationships among things and behaviors, in both small- and large-scale patterns. It does not give us ordinary reality, the world we live in *as* we live it. Instead of writing or talking exclusively *about* their experiences through abstract concepts, as one might do in applying productive scientific theory, trying to make language as invisible as possible while focusing on the objects of scientific expressions, poets report more concretely, *in* and *with* the facts and frameworks of what they see in themselves *in relation to* Others, in particular landscapes and emotional and social situations. They aim for representation from one self-conscious interiority to another in a manner that flags their language, stirs something up in us, finds the strange in the everyday, and takes us out of ourselves for a moment to show us something about ourselves in principle if not in precisely reported fact. Keeping people from jumping to culturally standardized conclusions about what is reported and evoking a larger sense of shared humanity in the process make the poetic effort both didactic and comparative.

The use of metaphor as a tool of and for discovery sets a similar instructional table. The idea that grandfather is an oak or that subatomic particles are best conceptualized as billiard balls in motion might prove to be the source for unraveling the very things that puzzle us most and the kinds of insights not available to us any other way. Knowing that, and consciously exploiting the “as if” world of analogies and creative equivalencies that is metaphor and that binds *all* interpretations of experience (for scientist and poet alike, albeit at different levels of disciplinary recognition), poets argue that the most productive form of inquiry depends on the nature of the problem to be addressed. Just as *Finnegan’s Wake* refuses reduction to the subject of free masonry or river journeys and resurrection or to critical summaries of its formal properties and

inspirations, so the statistical expression of cross-cultural trait distribution in aboriginal California cannot yield its exactness and rhetorical integrity to interpretive statements high in poeticity or uncommon metaphor. The implications of that are important, not only for helping us unravel the cultural constructions of a shared world that both differentiates and consolidates through various levels and circumstances, but also because changing the language of our descriptions, as Wittgenstein (1974) says, changes the analytic game itself, including changing the premises for research entry points. So poetics is not well conceived as a simple supplement to existing quantitative methods. It is a radically different form of interpreting, and therefore of knowing and reporting, whose domain is decidedly qualitative but no less complementary to other kinds of studies as a result.

Poetic inquiry can give us provocative and enlightening information about the nature of the world and our place in it, some of which is not available to the same extent, in the same form, or at all through other means. Ensuring interpretations grounded in self-awareness, it is also designed to keep premature closure on thinking in check while encouraging creativity in research and reporting. It is a way of reminding us that, despite numerous common denominators, science and poetics do different work; that no single genre or method can capture it all; that nothing we say can be nested in the entirely new; and that the field of experience and representation is by definition both culturally cluttered and incomplete for all of us at some level. Softening or solving such problems (e.g., by reaching beyond analytic categories whose only reality lies in the minds and agreements of the researchers themselves) matters if we are ever going to get a handle on the ordinary realities of the people we study—the universe *they* know, interpret, and act in as sentient beings. Such realities escape only at great cost to understanding ourselves, how we are articulated socially and semiotically, how we construct our Selves as meaningful entities in our own minds and in relation to each other, and what that contributes to acting responsibly in the shrinking space of a shared planet. For these and other reasons, privileging one form over the other as the source of Truth for all purposes is to confuse apples and hammers—to be satisfied with one tool for all jobs and with the politics of the moment. Inclusive is better. The combination of ethnological science and anthropological poetics can yield a more robust and complete accounting for and

representation of our existence as meaning-making and meaning-craving Beings-in-the-World.

—Ivan Brady

REFERENCES

- Brady, I. (2000). Anthropological poetics. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 949–979). Thousand Oaks, CA: Sage.
- Brady, I. (2003). *The time at Darwin's Reef: Poetic explorations in anthropology and history*. Walnut Creek, CA: AltaMira.
- Rothenberg, J., & Rothenberg, D. (Eds.). (1983). *Symposium of the whole: A range of discourse toward an ethnopoetics*. Berkeley: University of California Press.
- Tedlock, D. (1999). Poetry and ethnography: A dialogical approach. *Anthropology and Humanism*, 24(2), 155–167.
- Wittgenstein, L. (1974). *On certainty* (G. E. M. Anscombe & G. H. Von Wright, Eds.). Oxford, UK: Blackwell.

POINT ESTIMATE

A point estimate is the estimated value of a population PARAMETER from a SAMPLE. Point estimates are useful summary statistics and are ubiquitously reported in QUANTITATIVE analyses. Examples of point estimates include the values of such coefficients as the SLOPE and the INTERCEPT obtained from REGRESSION analysis. A point estimate can be contrasted with an *interval estimate*, also known as a CONFIDENCE INTERVAL. A confidence interval reflects the uncertainty, or ERROR, of the point estimate. A point estimate provides a single value for the population parameter, whereas a confidence interval provides a range of possible values for the population parameter. Although the true parameter cannot be known, the point estimate is the best approximation.

A point estimate facilitates description of the relationship between VARIABLES. The relationship may be STATISTICALLY SIGNIFICANT, but the magnitude of the relationship is also of interest. A point estimate provides more information about the relationship under examination. A hypothetical simple REGRESSION analysis is provided here to illustrate the concept of a point estimate. Given an INDEPENDENT VARIABLE X and a DEPENDENT VARIABLE Y in the regression equation $Y = a + bX + e$, let X = number of hours of sleep the night before a test, Y = exam grade (in the form of a percentage) in a sample of students in an American

politics seminar, b = slope, a = y -intercept, and e = error. An analysis of our hypothetical data produces the following regression equation: $\hat{Y} = -2 + 9.4X$. A point estimate is the value obtained for any sample statistic; here we are interested in the values of the intercept and the slope. The slope suggests that, on average, 1 additional hour of sleep is associated with an increase of 9.4 percentage points on the exam. If the slope estimate is also significant, the point estimate indicates that students who sleep more the night before a test perform substantially better. If, on the other hand, the slope estimate were very small—say, .5—the relationship would not appear to be substantively meaningful despite statistical significance. The example illustrates that point estimates provide valuable information about the relationship under examination.

The value of the intercept is also of interest. For this example, the value of the intercept is -2 . The intercept indicates that, on average, students with zero hours of sleep received a score of -2 percentage points; however, this value does not make sense because the grading scale ranges from 0 to 100. Also, if we were to plot these data along an x -axis and a y -axis, we would see that none or perhaps just a couple of students slept very little—say, 0 to 2 hours—the night before the test. With so few data points, PREDICTION of test scores for such little sleep should be avoided (in this case, we obtain a nonsensical result). Point estimates are useful for illustrating the relationship in more concrete terms and for making predictions but contain more error when based on relatively little data. In this case, the y -intercept underscores the caution that must be taken with point estimates.

The hypothetical example given here to illustrate the concept of a point estimate employed ORDINARY LEAST SQUARES as the ESTIMATOR; however, point estimates of parameters may be obtained using other estimators, such as MAXIMUM LIKELIHOOD. In fact, the point estimate, or the value obtained for the sample statistic, depends in part on the estimator chosen by the researcher. Thus, it is important for the researcher to select an appropriate or *preferred* estimator given the particular properties of the research question (Kennedy, 1998, pp. 5–6). Point estimates are also sensitive to the SPECIFICATION of the model, the presence of measurement error, the sample size, and the SAMPLE draw.

—Jacque L. Amoureux

REFERENCES

- Kennedy, P. (1998). *A guide to econometrics*. Cambridge: MIT Press.
- Lewis-Beck, M. S. (1980). *Applied regression: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–022). Newbury Park, CA: Sage.
- Lewis-Beck, M. S. (1995). *Data analysis: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–103). Thousand Oaks, CA: Sage.

POISSON DISTRIBUTION

The Poisson distribution measures the PROBABILITY that a DISCRETE positive variable takes on a given value for an observation period. For a given period, we observe the sum of “successes” among a possibly infinite number of trials. This sum has no upper bound. Trials or events are INDEPENDENT of one another and cannot occur simultaneously. In a given hour, for example, we might receive 10 phone calls from telemarketers; in this case, our Poisson variable takes on a value of 10. We define the EXPECTED VALUE ($E[Y]$) of a Poisson RANDOM variable as λ , or the anticipated rate of occurrence for the next period. Once having calculated a value for λ (drawing on results from previous periods), we can determine the probability with which Y takes on the value k using the Poisson PROBABILITY DENSITY FUNCTION:

$$\Pr(Y = k) = \frac{e^{-\lambda} \lambda^k}{k!}.$$

Although Siméon-Denis Poisson (1791–1840) originally developed the Poisson as an approximation for the binomial probability distribution function, scholars have since derived the Poisson from a simple set of propositions.

The Poisson bears a striking resemblance to the BINOMIAL DISTRIBUTION, which summarizes the outcome of n Bernoulli trials. Poisson derived the function as a way of approximating the probability that a binomial random variable takes on a particular value when the number of trials (n) is large, the chance of success (p) is small, and $\lambda = np$ is moderate in size—that is, in the case of rare events. (For a derivation, see Ross, 2002.) The variance formula for the Poisson follows from that for the binomial as well: $\text{Var}(Y) = np(1 - p) \approx \lambda$ when n is large and p

is small. The key difference between the binomial and the Poisson is that we know the number of trials (n) in a binomial distribution, whereas the number of trials in the Poisson is unknown and approaches infinity.

Social scientists have since begun to use the Poisson as a probability distribution in its own right to describe phenomena as diverse as the number of wars per year and the incidence of certain rare behaviors within a given population. Consider the case of the number of presidential vetoes during a given year. This example satisfies assumptions underlying the more general Poisson derivation (see King, 1998). Vetoes on particular bills are independent of one another. No two vetoes can occur at exactly the same time. Given a λ of 3.5 (the average of presidential veto rates during previous years), we can calculate the probability that the president would veto at least three bills in the next year (see Figure 1):

$$\begin{aligned} \Pr(k \geq 3) &= 1 - P(k = 0) - P(k = 1) - P(k = 2) \\ &= 1 - e^{-3.5} - 3.5e^{-3.5} - \frac{(3.5)^2}{2}e^{-3.5} \\ &= 1 - 10.625e^{-3.5} = 0.8641. \end{aligned}$$

Were we to vary the length of the interval, we would need an adjusted version of the equation that takes t (the length of the interval) as an additional PARAMETER. In this case, we would calculate instead the probability that the president will veto at least three bills during the next 2 years (see Figure 1):

$$\Pr(Y = k) = \frac{e^{-\lambda t} (\lambda t)^k}{k!}.$$

$$\begin{aligned} \Pr(k \geq 3) &= 1 - P(k = 0) - P(k = 1) - P(k = 2) \\ &= 1 - e^{-7} - 7e^{-7} - \frac{(7)^2}{2}e^{-7} \\ &= 1 - 32.5e^{-7} = 0.9704. \end{aligned}$$

Although the Poisson distribution as described above applies to single observations, social scientists have developed the POISSON REGRESSION technique to estimate λ for a set of observations.

—Alison E. Post

REFERENCES

- King, G. (1998). *Unifying political methodology*. Ann Arbor: University of Michigan Press.

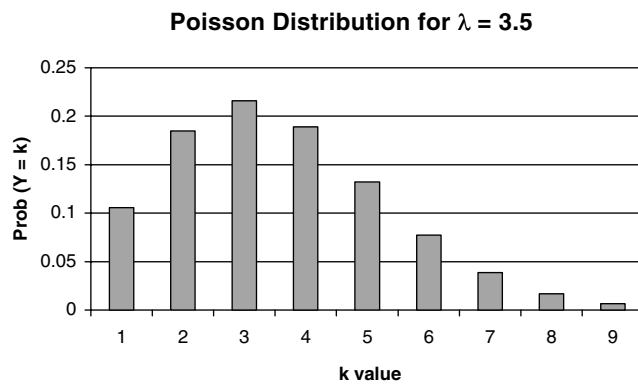


Figure 1 Poisson Distribution

Larsen, R. J., & Marx, M. L. (1990). *Statistics*. Englewood Cliffs, NJ: Prentice Hall.

Ross, S. (2002). *A first course in probability*. Upper Saddle River, NJ: Prentice Hall.

POISSON REGRESSION

A Poisson regression is a REGRESSION MODEL in which a dependent variable that consists of counts is modeled with a POISSON DISTRIBUTION. Count-dependent variables, which can take on only nonnegative integer values, appear in many social science contexts. For example, a sociologist might be interested in studying the factors that affect the number of times that an individual commits criminal behaviors. In this case, each individual in a data set would be associated with a number that specifies how many criminal behaviors he or she committed in a given time frame. Individuals who committed no such behaviors would receive a count value of 0, individuals who committed one crime would receive a count value of 1, and so forth. The sociologist in question could use a Poisson regression to determine which of a set of specified independent variables (e.g., an individual's level of educational attainment) has a significant relationship with an individual's criminal activity as measured by counts of criminal acts.

Poisson regressions are, as noted above, based on the Poisson distribution. This one-parameter univariate DISTRIBUTION has positive support over the nonnegative integers, and hence it is appropriate to use

such a distribution when studying a dependent variable that consists of counts.

MODEL SPECIFICATION

The MODEL SPECIFICATION for a Poisson regression is based on assuming that a count-dependent variable y_i has, conditional on a vector of independent variables $\mathbf{x}_i$, a Poisson distribution independent across other observations i with mean μ_i and the following probability DISTRIBUTION function:

$$f(y_i|\mathbf{x}_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}. \quad (1)$$

The link between μ_i and x_i is assumed to be

$$\mu_i = \exp(\mathbf{x}_i' \boldsymbol{\beta}), \quad (2)$$

where $\boldsymbol{\beta}$ is a parameter vector (with the same dimension as $\mathbf{x}_i$) to be estimated. Note that the exponential link function between μ_i and $\mathbf{x}_i$ ensures that μ_i is strictly positive; this is essential because the mean of the Poisson distribution cannot be nonpositive. Based on a set of n observations $(y_i, \mathbf{x}_i)$, equations (1) and (2) together generate a log-likelihood function of

$$\log L = \sum_{i=1}^n [y_i \mathbf{x}_i' \boldsymbol{\beta} - \exp(\mathbf{x}_i' \boldsymbol{\beta}) - \log(y_i!)], \quad (3)$$

where $\log(\cdot)$ denotes natural LOGARITHM. The first-order conditions from this log-likelihood function—there is one such condition per each element of $\boldsymbol{\beta}$ —are NONLINEAR in $\boldsymbol{\beta}$ and cannot be solved in closed form. Therefore, numerical methods must be used to solve for the value of $\hat{\boldsymbol{\beta}}$ that maximizes equation (3).

Interpretation of Poisson regression output is similar to interpretation of other nonlinear regression models. If a given element of $\hat{\boldsymbol{\beta}}$ is positive, then it follows that increases in the corresponding element of $\mathbf{x}_i$ are associated with relatively large count values in y_i . However, comparative statics in a Poisson regression model are complicated by the fact that the derivative of μ_i with respect to an element of $\mathbf{x}_i$ in equation (2) depends on all the elements of $\mathbf{x}_i$ (cross-partial derivatives of μ_i with respect to different elements of $\mathbf{x}_i$ are not zero, as they are in an ORDINARY LEAST SQUARES regression). Hence, with a vector estimate $\hat{\boldsymbol{\beta}}$, one can vary an element of $\mathbf{x}_i$ and, substituting $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ in equation (2), generate a sequence of estimated μ_i values based on

a modified $\mathbf{x}_i$ vector. It is then straightforward to generate a probability distribution over the nonnegative integers—an estimated distribution for y_i —using the link function in equation (2). This distribution, it is important to point out, will be conditional on the elements of $\mathbf{x}_i$ that are not varied.

The log likelihood in equation (3) satisfies standard regularity conditions, and from this it follows that inference on $\hat{\beta}$ can be carried out using standard asymptotic theory. Accordingly, significance tests, likelihood ratio tests, and so forth can easily be calculated based on $\hat{\beta}$ and an estimated covariance matrix for this parameter vector. The Poisson regression model is a special case of a GENERALIZED LINEAR MODEL (McCullagh & Nelder, 1989); hence, generalized linear model theory can be applied to Poisson regressions.

EXTENSIONS

The standard Poisson regression model—encapsulated in equations (1) and (2)—can be extended in a number of ways, many of which are described in Cameron and Trivedi (1998), a detailed and very broad reference on count regressions. One very common extension of the Poisson model is intended to reflect the fact that a count-dependent variable y_i can, in practice, have greater variance than it should based on the nominal Poisson variance conditional on a vector $\mathbf{x}_i$. Note that a Poisson random variable with mean λ has variance λ as well. This means that modeling the mean of y_i as in equation (2) imposes a strong constraint on the variance of y_i as well.

When, given observations $(y_i, \mathbf{x}_i)$, the variance of y_i exceeds μ_i because of unmodeled HETEROGENEITY, then y_i is said to suffer from overdispersion (i.e., it has greater variance than one would expect given observed y_i and $\mathbf{x}_i$). An overdispersed count-dependent variable can be studied using a negative binomial regression model; the standard Poisson model should not be applied to overdispersed data as results based on this model can be misleading.

Moreover, the Poisson regression model is a special case of the negative binomial model. Thus, when applying a negative binomial regression model to a given count data set, it is straightforward to test whether the data set, conditional on independent variables in the associated $\mathbf{x}_i$ vector, is overdispersed, implying that a standard Poisson model should not be used to analyze it.

Another extension of the standard Poisson model, due to Lambert (1992), is intended to model a count-dependent variable with an excessive number of zero observations (i.e., an excessive number of values of y_i such that $y_i = 0$). Such a set of dependent variable values can be labeled zero inflated, and Lambert's zero-inflated Poisson regression model assumes that some observations y_i are drawn from a standard Poisson regression model (or a negative binomial model) and that some values of y_i are from an "always zero" model. Consequently, associated with each y_i that is observed zero is a latent probability of being drawn from the always zero model and, correspondingly, a latent probability of being drawn from a regular Poisson model that puts positive weight on zero and also on all positive integers. A standard Poisson regression is not a special case of Lambert's zero-inflated Poisson model, and as pointed out in Greene's (2000) discussion and survey of count regressions, testing for excess zeros requires nonnested hypothesis testing.

Additional extensions of the Poisson regression model can be found in Cameron and Trivedi (1998) and in Greene (2000). Cameron and Trivedi, for example, discuss TIME-SERIES extensions of the Poisson model as well as generalizations that allow the modeling of vectors of counts.

—Michael C. Herron

REFERENCES

- Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. New York: Cambridge University Press.
- Greene, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics*, 34, 1–14.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

POLICY-ORIENTED RESEARCH

Policy-oriented research is designed to *inform* or *understand* one or more aspects of the public and social policy process, including decision making and policy formulation, implementation, and evaluation. A distinction may be made between research *for* policy and research *of* policy. Research *for* policy is research that informs the various stages of the policy

process (before the formation of policy through to the implementation of policy). Research *of* policy is concerned with how problems are defined, agendas are set, policy is formulated, decisions are made, and how policy is implemented, evaluated, and changed (Nutley & Webb, 2000, p. 15). Policy-oriented research (POR) can simultaneously be of both types.

How POR actually informs policy (if at all) is a debated issue. Young, Ashby, Boaz, and Grayson (2002) identify various models that help us understand this process. A *knowledge-driven model* assumes that research (conducted by experts) *leads* policy. A *problem-solving model* assumes that research *follows* policy and that policy issues shape research priorities. An *interactive model* portrays research and policy as mutually influential. A *political/tactical model* sees policy as the outcome of a political process, where the research agenda is politically driven. An *enlightenment model* sees research as serving policy in indirect ways, addressing the context within which decisions are made and providing a broader frame for understanding and explaining policy. Each model can help us understand the links between POR and specific policy developments. Young and colleagues (2002, p. 218) suggest that the role of policy research “is less one of problem solving than of clarifying issues and informing the wider public debate.”

Thus, POR may serve several functions and can have a diverse range of audiences. It provides a *specialist* function of informing and influencing the policy process and understanding how policy works and what works for target audiences of policy makers, policy networks and communities, research-aware practitioners, and academics; it also provides a *democratic* function, in which research findings contribute to the development of an informed and knowledge-based society as well as to the broader democratic process. Here the users of research will include organized groups with vested interests, service users, and the public as a whole.

Policymakers and those who implement policy (e.g., social workers, health care workers) are increasingly looking to base their policy and practice on “evidence,” particularly on evidence of effectiveness and “what works.” What counts as evidence (and who says so) are highly contested areas. Although POR can be a key source of “evidence,” what actually counts as evidence—especially in informing policy and practice—will be far broader than research findings alone.

POR can draw from the full range of possible research designs and methods (Becker & Bryman, 2004). Depending on the specific research question(s) to be addressed, in some cases, just one method will be used. In other cases, there may be an integration of different methods either within a single piece of research or as part of a wider program of research being conducted across multiple sites or cross-nationally. Although each method and design has its own strengths and limitations, RANDOMIZED CONTROLLED TRIALS, followed by SYSTEMATIC REVIEWS and statistical META-ANALYSIS, are seen by many as the “gold standards.” This is particularly the case in medical and health policy research, and there is an ongoing debate concerning which research designs and methods are the most appropriate for POR in social care, education, criminal justice, and other fields of public and social policy.

—Saul Becker

REFERENCES

- Becker, S., & Bryman, A. (Eds.). (2004). *Understanding research for social policy and practice: Themes, methods and approaches*. Bristol, UK: Policy Press.
- Nutley, S., & Webb, J. (2000). Evidence and the policy process. In H. Davies, S. Nutley, & P. Smith (Eds.), *What works? Evidence-based policy and practice in public services* (pp. 13–41). Bristol, UK: Policy Press.
- Young, K., Ashby, D., Boaz, A., & Grayson, L. (2002). Social science and the evidence-based policy movement. *Social Policy and Society*, 1(3), 215–224.

POLITICS OF RESEARCH

With politics in relation to research, the issue is how far it can be or should seek to be value free. The early and traditional argument was that social research should strive to be scientific and OBJECTIVE by emulating the practices and assumptions of the natural sciences, a philosophical position known as POSITIVISM. The key characteristics focused on using a value-free method, making logical arguments, and testing ideas against that which is observable to produce social facts. Such principles have always been subject to dispute. Nevertheless, they dominated mainstream thought during the 19th century and a considerable part of the 20th century.

More recently, alternative positions have prevailed. Positivism has been challenged as inappropriate to the

social sciences, which is concerned with social actors' meanings, beliefs, and values. The reliance on quantitative methods, such as SURVEYS, has been modified as greater legitimacy is afforded to in-depth qualitative studies. There has been a blurring of the boundaries between social research and policy and practice (Hammersley, 2000). As a result, the assumption that social science can produce objective knowledge has been challenged.

Hammersley (1995) has argued that it is necessary to ask three different questions as to whether social research is political. *Can* research be nonpolitical? Does research as *practiced* tend to be political? *Should* research be political? Some social scientists argue that it is impossible for research not to involve a political dimension. All research is, necessarily, influenced by social factors, such as how the topic is framed, what access to participants is possible, and the role of the researcher in the research process. Funding bodies also have a major impact in terms of which projects they finance and the conditions relating to this.

Others also claim that it is neither possible nor desirable for social scientists to adopt a disinterested attitude to the social world. Sociology, for instance, has always been driven by a concern for the "underdog" and by its willingness to challenge the status quo. Currently, feminist and antiracist research is explicitly aimed at bringing about change. However, there are still those who argue that research should not be directly concerned with any goal other than the production of knowledge. From this view, research has a limited capacity to deliver major political transformations, and such pursuits limit the value of research itself (Hammersley, 1995).

A possible way out of this conundrum is to distinguish between politics, on one hand, and objectivity in producing knowledge, on the other. Asking questions that are guided by a political or ethical point of view need not prevent a researcher from striving toward objectivity in the actual research. This involves using rigorous and systematic procedures throughout the research process, including during the analysis and when producing an account. Researchers are arguing that, even if there is no utopia of absolute objectivity for social scientists, this must still be an aspiration in the interest of generating robust knowledge and understandings (Christians, 2000).

—Mary Maynard

REFERENCES

- Christians, C. (2000). Ethics and politics in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The handbook of qualitative research* (pp. 133–155). Thousand Oaks, CA: Sage.
- Hammersley, M. (1995). *The politics of social research*. London: Sage.
- Hammersley, M. (2000). *Taking sides in social research*. London: Routledge.

POLYGON

Polygon is a type of chart that displays a variable of ordinal, interval, or grouped ratio scale on the x -axis and another variable of interest, most often frequencies of observations, on the y -axis.

—Tim Futing Liao

See also FREQUENCY POLYGON

POLYNOMIAL EQUATION

When an equation contains right-side variables that are exponentiated to some power greater than 1, we call it a *polynomial equation*. Polynomial equations are observed in research in economics more frequently than in other social sciences. For example, the theory of marginal cost suggests a U-shaped relation between output and marginal cost. As we increase output of a commodity, the marginal cost for producing one unit of the commodity decreases due to the economy of scale. As we keep production beyond a certain amount of output, however, the cost tends to increase due to saturation of the production facility. Mathematically, this relationship can be captured with a quadratic equation of the form $\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$.

STOCHASTIC FORM OF POLYNOMIAL EQUATIONS

Statistical specifications sometimes represent the theory associated with polynomial equations in a STOCHASTIC form. For example, the stochastic version of the preceding equation is $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$. If we expand our discussion to a more general case in which there are more than two powers for several

variables, we have the following general expression of the stochastic polynomial equation:

$$Y = \beta_0 + \sum_{i=1}^m \beta_i X^i + \cdots + \sum_{j=1}^n \zeta_j Z^j + \varepsilon.$$

In the remainder of this entry, we shall consider polynomial equations in a stochastic context.

For example, polynomial specification of a REGRESSION model is a special case of a NONLINEAR model, which can be represented more generally as

$$\begin{aligned} Y &= \alpha g_1(\mathbf{z}) + \beta g_2(\mathbf{z}) + \cdots + \zeta g_k(\mathbf{z}) + \varepsilon \\ &= \alpha X_1 + \beta X_2 + \cdots + \zeta X_k + \varepsilon. \end{aligned}$$

Here a vector $\mathbf{z}$ consists of several independent variables. A series of TRANSFORMATION functions, such as $g_1(\mathbf{z})$, $g_2(\mathbf{z})$, $g_k(\mathbf{z})$, specify various nonlinear relations among the variables in $\mathbf{z}$. Notice that we begin with a nonlinear specification of independent variables but end up with a familiar expression of the right side of the classic linear regression equation, $\alpha X_1 + \beta X_2 + \cdots + \zeta X_k + \varepsilon$, through the transformation of the independent variables. This shows the intrinsic linearity of the model. If we find the nature of the nonlinear relationship, we can often transform our complex model into a simple linear form. A polynomial equation is one of the possible specifications of the transformation function. LOGARITHMIC, exponential, and reciprocal equations can also be alternatives, depending on the nature of the nonlinearity.

ESTIMATION OF POLYNOMIAL REGRESSIONS

We can estimate a polynomial regression by classic ORDINARY LEAST SQUARES. This is because nonlinearity in independent variables does not violate the assumption of linearity in parameters. We can just put the transformed values of independent variables into the equation. Then, we get a simplified form of LINEAR REGRESSION. Polynomial terms are often associated with MULTICOLLINEARITY because variables in quadratic and cubic forms will be correlated. However, estimates from polynomial regression equations are best linear unbiased estimates (BLUE) if all of the regression assumptions are met.

DETECTION OF POLYNOMIAL RELATIONSHIPS

Given this basic definition and characteristics of the polynomial regression equation, a natural question is how we can specify and detect polynomial relationships. There are two approaches to answering this

question. The first is grounded in theory, and the second is grounded in statistical testing. Based on a well-developed theory of the phenomenon we are studying, such as the marginal cost theory mentioned above, we could specify a priori predicted relations among variables and then just check the STATISTICAL SIGNIFICANCE of the coefficients. Obviously, this test of significance is a natural test of the theory. However, in many cases, theory will be too weak to specify particular polynomial or nonlinear forms.

Thus, we can also address the problem of specifying and detecting polynomial relationships statistically. Three approaches are commonly used. First, we can analyze the residuals from a simple linear regression graphically. Let's think of a simple polynomial equation with one variable. Suppose the true model is $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$, but we estimated it with a sample regression of $Y = \hat{\beta}_0 + \hat{\beta}_1 X + \varepsilon$. Then, the residuals will reflect $\beta_2 X^2$ as well as ε . In other words, the residuals will contain systematic variations of the true model in addition to the pure disturbances. If we rearrange the observations in ascending order of the x values, then we will easily find a pattern in the residuals.

Second, we can run piecewise linear regressions. Taking the same example of the residual analysis, we can rearrange the observations again in ascending order and divide the sample into several subgroups. If we run separate regressions on the subgroups, we will have coefficient estimates that are consistently increasing or decreasing depending on whether the sign of β_2 is positive or negative.

Third, if we deal with the problem of specification and detection in the more general context of nonlinearity, then we can use a Box-Cox specification. According to the Box-Cox transformation, we specify a nonlinear functional form of the independent variables as

$$X = \frac{Z^\lambda - 1}{\lambda}$$

Here λ is an unknown parameter, so we have a variety of possible functional forms depending on the value of λ . Critical to our current interest is the case of $\lambda = 1$, where we have X as a linear function of Z . A LIKELIHOOD RATIO test is often used to see whether the likelihood ratio statistics that are estimated with and without the restriction of $\lambda = 1$ suggest a difference between the two specifications. If we find a statistically significant difference between the two, we reject the null hypothesis of linearity in variables. Given the results of these statistical procedures, we can make a judgment on the existence of nonlinearity. However,

the existence of nonlinearity itself does not tell us about the true nature and form of nonlinearity. In a general specification of nonlinearity, we can use the Box-Cox transformation equation for estimating the value of λ . The value of λ is indicative of the nature of the nonlinearity. However, the Box-Cox approach does not cover all possible forms of nonlinear relations. Moreover, the Box-Cox approach does not pertain specifically to polynomial-based nonlinearity.

CONCLUSION

In summary, specification and detection of nonlinearity, including polynomial forms, should be grounded in theory and confirmed through statistical testing. A purely theoretical specification may miss the exact form of nonlinearity. A purely statistical basis for specifying nonlinearity is likely to provide a weak basis for using nonlinear forms. Thus, considerable judgment is required in specifying polynomial and other forms of nonlinearity.

—B. Dan Wood and Sung Ho Park

REFERENCES

- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Mahwah, NJ: Prentice Hall.
- Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.

POLYTOMOUS VARIABLE

A polytomous variable is a variable with more than two distinct categories, in contrast to a DICHOTOMOUS variable. In principle, any CONTINUOUS variable or any INTERVAL or RATIO variable with more than two observed values is a polytomous variable, but the term is usually reserved for NOMINAL and ORDINAL variables, called unordered and ordered polytomous variables. Usually, when applied to ordinal variables, the term is used to refer to ordinal variables with relatively few (e.g., less than 10) distinct categories; ordinal variables with large numbers of categories (e.g., more than 20) are sometimes described as, and often treated the same as, continuous interval or ratio variables. Also, although in principle the term is applicable to either INDEPENDENT or DEPENDENT VARIABLES in any analysis, the term is usually reserved for dependent variables in LOGIT and LOGISTIC

REGRESSION models. Other terms sometimes used for polytomous variables are *polychotomous* (just an extra syllable, also in contrast to dichotomous variables) and *multinomial* (in contrast to binomial). Because the terms *binomial* and *multinomial* refer to specific probability distributions and to variables having those distributions, the term *polytomous* may be preferred when no implication about the distribution of the variable is intended.

—Scott Menard

POOLED SURVEYS

Pooled surveys represent the combining of different SURVEYS into one data set. Usually, the surveys are essentially the same measuring instruments, applied repeatedly over time. For example, the analyst might combine, or pool, the presidential election year surveys of the American National Election Study from, say, 1952 to 2000. That would mean a pool of 13 cross-sectional surveys conducted across a time period of 48 years. Advantages of the pooled survey are a great increase in SAMPLE size and VARIANCE. Analysis complications can arise from the mix of TIME-SERIES and CROSS-SECTIONAL DESIGNS. Also, the study may be forced to restrict itself to the narrower range of items that were posed in all the surveys.

—Michael S. Lewis-Beck

POPULATION

A population is the universe of units of analysis, such as individuals, social groups, organizations, or social artifacts about which a researcher makes conclusions. In the course of the study, the researcher collects measurable DATA about a specific population in which he or she is interested. In a study, the units of analysis that comprise the population are referred to as subjects. Although the goal of all studies is to make inferences about population, in most cases, it is virtually impossible to gather data about all members of the population in which the researcher is interested. Because some populations, such as college students or Native Americans, are very large, attempts at complete enumeration of all cases are extremely time-consuming, prohibitively expensive, and prone to data

collection error. Selecting a representative subset of the population, called a **SAMPLE**, is often preferable to enumeration. By using inferential statistical methods, the researcher is able to make conclusions about characteristics of a population, based on a **RANDOM SAMPLE** of that population.

Some social studies focus on populations that actually exist, such as freshman students of the New York University. Alternatively, a study can be focused on a conceptual population that does not exist in reality but can be hypothetically conceptualized, such as divorced middle-class males age 45 and older.

Characteristics of the population are generally referred to as **PARAMETERS**. By convention, parameters are denoted by Greek letters, such as μ for population mean and σ for population standard deviation. Population parameters are fixed values and are, in most cases, unknown. For example, the percentage of people in favor of abortions in Utah may be unknown, but theoretically, if several researchers measured this parameter at a specific time, they would come up with the same number. In contrast, sample statistics vary from one sample to another. It would be extremely unlikely for several random samples measuring the percentage of pro-choice Utah residents to arrive at the same number. In contrast to population parameters, the values of statistics for a particular sample are either known or can be calculated. What is often unknown is how representative the sample actually is of the population from which it was drawn or how accurately the statistics of this sample represent population parameters.

Prior to conducting the actual study, the researcher usually makes **ASSUMPTIONS** about the population to which he or she is generalizing and the **SAMPLING** procedures he or she uses for the study. These assumptions usually fall into one of two categories: (a) assumptions of which the researcher is virtually certain or is willing to accept and (b) assumptions that are subject to debate and empirical proof and in which the researcher is therefore most interested. Assumptions in the first category are called the **MODEL**. The assumptions in the second category are called the **HYPOTHESIS**. In the course of the study, after the assumptions about the unknown population parameters are made, the researcher tests the hypothesis and tries to determine how likely his or her sample statistics would be if these assumptions were actually true.

—Alexei Pavlichev

REFERENCES

- Agresti, A., & Finlay, B. (1997). *Statistical methods for social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Babbie, E. R. (1995). *The practice of social research* (7th ed.). Belmont, CA: Wadsworth.
- Blalock, H. M., Jr. (1979). *Social statistics* (Rev. 2nd ed.). New York: McGraw-Hill.

POPULATION PYRAMID

The age-sex composition of a **POPULATION** is often portrayed in a stylized graph called a *population pyramid*, which shows the numbers of people, expressed in absolute numbers, percentages, or proportions, in each age (or year of birth) interval arrayed by sex. The name derives from the typical triangular or pyramidal shape of the image for populations that are growing because of high levels of fertility sustained over a substantial period of time. Pyramids for populations that are growing very slowly or are stable in size tend to be shaped like beehives, whereas pyramids for populations that are decreasing in size because of low rates of fertility show marked constrictions at the base.

The merits of population pyramids lie in their clear presentation of the relative sizes of age (or year of birth) cohorts and thus the relative sizes of important age-bounded groups such as the potential labor force or dependent populations such as children and the elderly. Population pyramids can also show the impact of trends and fluctuations in historic and recent levels of vital rates, especially fertility rates, on the population's age structure. Because population pyramids show both sexes, they also show the impact of the sex ratio at birth, which approximates 105 males to 100 females in most national populations, and the impact of sex-specific differentials in mortality and migration. Population pyramids also sometimes indicate patterns of error in age-specific counts. It is becoming very common for national statistical agencies to publish the age-sex tabulations necessary for constructing population pyramids on the Web.

The population pyramid displayed in Figure 1 shows the age-sex structure of the population of the Republic of Ecuador in 1990, with age reported in single years. The graph shows the typical features of a population pyramid for a developing country. Its overall shape is triangular, the result of historically high levels of fertility. It is asymmetric with respect to sex: The base

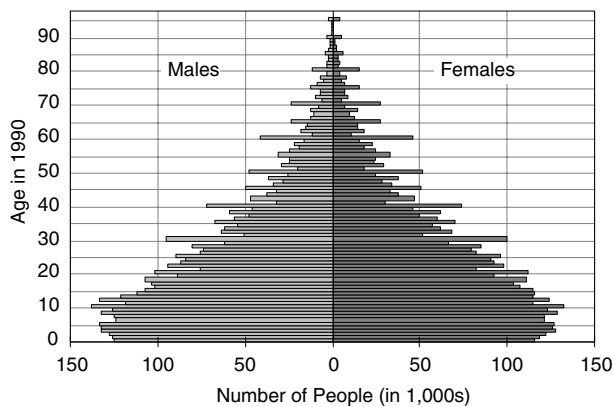


Figure 1 Population Pyramid for Ecuador, 1990

SOURCE: Instituto Nacional de Estadística y Censos, 1990, www4.inec.gov.ed/poblac/pobla03.txt.

of the pyramid has a slight preponderance of males because of the male-biased sex ratio among births, and the top has a slight preponderance of females because mortality rates slightly favor females. There is also clear evidence of error in the age-specific counts due to “age heaping,” a common pattern of age misreporting in which people are more likely to report ages ending in selected digits (often 0 and 5, as observed here) than ages ending in other digits. (This type of error is often masked by the common practice of aggregating age or year of birth categories into 5-year intervals.) The inward cut at the bottom of the pyramid could reflect either the underenumeration of children in the census or a decline in fertility rates in Ecuador during the 1980s. Comparing the sizes of the cohorts ages 0 to 9 enumerated in the 1990 census with the sizes of the cohorts ages 10 to 19 enumerated in the 2000 census suggest that the constriction at the base of the pyramid is mostly attributable to underenumeration of children in the 1990 census.

—Gillian Stevens

POSITIVISM

Positivism is a philosophy of science that rejects metaphysical speculation in favor of systematic observation using the human senses. “Positive” knowledge of the world is based on GENERALIZATIONS from such observations that, given sufficient number and

consistency, are regarded as producing laws of how phenomena coexist or occur in sequences. In the 19th century, positivism was not merely a philosophy of science; it expressed a more general worldview that lauded the achievements of science.

CENTRAL TENETS

Although numerous attempts have since been made to identify the central tenets of positivism, the following are generally accepted as its main characteristics.

1. *Phenomenalism.* Knowledge must be based on experience, on what observers can perceive with their senses. This perception must be achieved without the subjective activity of cognitive processes; it must be “pure experience” with an empty consciousness.

2. *Nominalism.* Any abstract concepts used in scientific explanation must also be derived from experience; metaphysical notions about which it is not possible to make any observations have no legitimate existence except as names or words. Hence, the language used to describe observations must be uncontaminated by any theoretical notions.

3. *Atomism.* The objects of experience, of observation, are regarded as discrete, independent atomic impressions of events, which constitute the ultimate and fundamental elements of the world. Insofar as these atomic impressions are formed into generalizations, they do not refer to abstract objects in the world, only regularities among discrete events.

4. *General laws.* Scientific theories are regarded as a set of highly general lawlike statements; establishing such general laws is the aim of science. These laws summarize observations by specifying simple relations or constant conjunctions between phenomena. EXPLANATION is achieved by subsuming individual cases under appropriate laws. These laws are general in scope, in that they cover a broad range of observations, and are universal in form, in that they apply, without exception, across time and space.

5. *Value judgments and normative statements.* “Facts” and “values” must be separated as values do not have the status of knowledge. Value statements have no empirical content that would make them susceptible to any tests of their validity based on observations.

6. *Verification.* The truth or falsity of any scientific statement can be settled with reference to an observable

state of affairs. Scientific laws are verified by the accumulation of confirming evidence.

7. *Causation*. There is no CAUSALITY in nature, only regularities or constant conjunctions between events, such that events of one kind are followed by events of another kind. Therefore, if all we have are regularities between types of events, then explanation is nothing more than locating an event within a wider ranging regularity.

VARIETIES OF POSITIVISM

According to Halfpenny (1982), it is possible to identify 12 varieties of positivism. However, following Outhwaite (1987), these can be reduced to 3. The first was formulated by August Comte (1798–1857) as an alternative to theological and metaphysical ways of understanding the world. He regarded all sciences as forming a unified hierarchy of related levels, building on mathematics at the lowest level, followed by astronomy, physics, chemistry, biology, and, finally, sociology.

The second version, known as *logical positivism*, was founded in Vienna in the 1920s. The catch-cry of these philosophers was that any concept or proposition that does not correspond to some state of affairs (i.e., cannot be verified by experience) is regarded as being meaningless (*phenomenalism*). At the same time, it is argued that the concepts and propositions of the higher level sciences *can* be reduced to those of the lower ones. In other words, they adopted the reductionist position that the propositions of the social sciences could ultimately be analyzed down to those of physics.

The third version, which was derived from the second and is sometimes referred to as the “standard view” in the philosophy of science, dominated the English-speaking world in the post–World War II period. Its fundamental tenet is that all sciences, including the social sciences, are concerned with developing explanations in the form of universal laws or generalizations. Any phenomenon is explained by demonstrating that it is a specific case of some such law. These laws refer to “constant conjunctions” between events or, in the case of the social sciences, statistical correlations or regularities (“general law”).

This third version is based on the belief that there can be a *natural* scientific study of people and society, the doctrine known as the *unity of scientific method*. It is argued that despite the differences in subject matter of the various scientific disciplines, both natural and

social, the same method or logic of explanation can be used, although each science must elaborate these in a way appropriate to its objects of enquiry.

At its most general, positivism is a theory of the nature, omniscience and unity of science. In its most radical shape it stipulates that the only valid kind of (non-analytic) knowledge is scientific, that such knowledge consists in the description of the invariant patterns, the co-existence in space and succession over time, of observable phenomena. . . . Its naturalistic insistence on the unity of science and scientific disavowal of any knowledge apart from science induce its aversion to metaphysics, insistence upon a strict value/fact dichotomy and tendency to historicist confidence in the inevitability of scientifically mediated progress. (Bhaskar, 1986, p. 226)

POSITIVISM IN THE SOCIAL SCIENCES

It was through the work of August Comte and Emile Durkheim that positivism was introduced into the social sciences. Forms of positivism have dominated sociology, particularly in the decades immediately following World War II, and continue to do so today in disciplines such as psychology and economics. However, in recent years, positivism has been subjected to devastating criticism (for reviews, see Blaikie, 1993, pp. 101–104; Bryant, 1985).

CRITICISM

Some of the main points of dispute are the claim that experience can be a sound basis for scientific knowledge, that science should deal only with observable phenomena and not with abstract or hypothetical entities, that it is possible to distinguish between an atheoretical observation language and a theoretical language, that theoretical concepts have a 1:1 correspondence with “reality” as it is observed, that scientific laws are based on constant conjunctions between events in the world, and that “facts” and “values” can be separated.

Positivism has been attacked from many perspectives, including INTERPRETIVISM, FALSIFICATIONISM, and CRITICAL REALISM. The interpretive critique has focused on positivism’s inadequate view of the nature of social reality—its ONTOLOGY. Positivism

takes for granted the socially constructed world that interpretivism regards as social reality. Positivists construct fictitious social worlds out of the meaning it has for *them* and neglect what it means to the social actors.

The central feature of the falsificationist critique is positivism's process for "discovering" knowledge and the basis for justifying this knowledge. First, because they regard experience as an inadequate source of knowledge, and as all observation involves interpretation, falsificationists have argued that it is not possible to distinguish between observational statements and theoretical statements; all statements about the world are theoretical, at least to some degree. Second, it is argued that experience is an inadequate basis for justifying knowledge because it leads to a circular argument. On what basis can experience be established as a justification for knowledge except by reference to experience?

Positivism's claim that reality can be perceived directly by the use of the human senses has been thoroughly discredited. Even if it is assumed that a single, unique physical world exists independently of observers—and this assumption is not universally accepted—the process of observing it involves both conscious and unconscious interpretation. Observations are "theory laden." The processes that human beings use to observe the world around them, be it in everyday life or for scientific purposes, are not analogous to taking photographs. In "reading" what impinges on our senses, we have to engage in a complex process that entails both the use of concepts peculiar to the language of a particular culture and expectations about what is "there." Furthermore, we do not observe as isolated individuals but as members of cultural or subcultural groups that provide us with ontological assumptions. Therefore, observers are active agents, not passive receptacles.

The realist solution to the theory-laden nature of observation and description is to draw the distinction between the transitive and intransitive objects of science. Although our descriptions of the *empirical* domain may be theory dependent, the structures and mechanisms of the *real* domain exist independently of our descriptions of them. Reality is not there to be observed as in positivism, nor is it just a social construction; it is just there. Therefore, according to the realists, the relative success of competing theories to represent this reality can be settled as a matter of rational judgment (see REALISM and CRITICAL REALISM).

As well as accusing positivism of having an inadequate ontology, critical realism also attacked the positivist method of explanation in terms of constant conjunctions between events. Even if two kinds of phenomena can be shown to occur regularly together, there is still the question of why this is the case. According to critical realism, establishing regularities between observed events is only the starting point in the process of scientific discovery. Constant conjunctions of events occur only in closed systems, those produced by experimental conditions. However, positivism treats the world of nature as a closed system. In open systems characteristic of both nature and society, a large number of generative mechanisms will be exercising their powers to cause effects at the same time. What is observed as an "empirical" conjunction may, therefore, not reflect the complexity of the mechanisms that are operating.

—Norman Blaikie

REFERENCES

- Bhaskar, R. (1986). *Scientific realism and human emancipation*. London: Verso.
- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Bryant, C. G. A. (1985). *Positivism in social theory and research*. London: Macmillan.
- Halfpenny, P. (1982). *Positivism and sociology: Explaining social life*. London: Allen & Unwin.
- Outhwaite, W. (1987). *New philosophies of social science: Realism, hermeneutics and critical theory*. London: Macmillan.

POST HOC COMPARISON. See MULTIPLE COMPARISONS

POSTEMPIRICISM

EMPIRICISM has long appeared synonymous with science. After all, if science could not gain an insight into reality, what could? In the history of empiricism, it was John Locke who argued that our understanding of the world arises from our experiences. Nevertheless, he also emphasized that classification must be based on a view of the essential qualities of objects. Therefore,

there is no privileged access to things in themselves but to their properties (color, shape, feel, etc.). The empiricist philosopher David Hume was even more of a skeptic and held that only appearances can be known and nothing beyond them.

In what is a complicated history, EMPIRICISM provided the basis for a unity of method in the sciences. It also provided for a strict demarcation between the social and natural sciences centered on, for example, such issues as the neutrality of the process of observation, the unproblematic nature of experience, a strict separation of DATA from THEORY, and a universal approach to the basis of knowledge. In their different ways, Thomas Kuhn, Paul Feyerabend, and Mary Hesse added to postempiricist insights that take aim at these idealist assumptions and, in so doing, open up science to a more social and historical understanding that is rooted in EXPLANATIONS and understandings of the conditions that inform knowledge production, development, and application.

To simplify his argument, Kuhn (1970), in *The Structure of Scientific Revolutions*, characterized science as consisting of periods of “normal science” in which “puzzle solving” took place against the background of particular PARADIGMS. These comprise the intellectual standards and practices of a scientific community and are based on shared sets of assumptions. From time to time, however, anomalies are bound to occur, and key theories will seem to be falsified. When this happens and key scientists begin to challenge the orthodoxy, crises will occur. This will then spread, with a resulting revolution that leads to a new paradigm being established and so on.

A number of implications followed from Kuhn's (1970) groundbreaking work and the work of the postempiricists. First, the LOGICAL POSITIVISTS had held that there must be a strict separation between the language of theory and the language of OBSERVATION. Although Karl Popper (1959) has submitted this to critique, Kuhn took it a step further. The simple demarcation between the social and natural sciences could no longer hold as one between facts and values. The anti-essentialism of the postempiricists met with their commitment to detailed analysis based on the social context of production as a practical activity rooted within particular circumstances. Therefore, the social was not a separate activity to scientific endeavor but was fundamentally a part of it. Paul Feyerabend (1978) took this even further than the conservative orientation of Kuhn and held not only that local factors determined

how science progressed but, because of this, that there could be no grounds on which any rule might inform the practice of science.

The postempiricist critiques opened up a rich tradition of social studies of science. It has also left a legacy around a central problem neatly expressed by James Bohman (1991): “how to derive standards for the comparison and evaluation of competing explanations and research programs without falling back into ahistorical and essentialist epistemology” (p. 4).

—Tim May

REFERENCES

- Bohman, J. (1991). *New philosophy of social science: Problems of indeterminacy*. Cambridge, UK: Polity.
- Feyerabend, P. (1978). *Against method*. London: Verso.
- Kuhn, T. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.
- Williams, M., & May, T. (1996). *Introduction to the philosophy of social research*. London: Routledge Kegan Paul.

POSTERIOR DISTRIBUTION

A posterior distribution $p(\theta|\mathbf{x})$ is a probability distribution describing one's subjective belief about an unknown quantity θ after observing a data set $\mathbf{x}$. Posterior distributions are constructed using Bayes' rule and occupy a central position in BAYESIAN INFERENCE. Denote the PRIOR DISTRIBUTION by $p(\theta)$ and the likelihood function by $p(\mathbf{x}|\theta)$. Bayes' rule is often written as

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)p(\theta).$$

Bayes' rule describes the process of learning from data. It explains how two people with different prior beliefs can be drawn toward similar conclusions after observing $\mathbf{x}$. The proportionality in Bayes' rule is resolved by the fact that $p(\theta|\mathbf{x})$ is a probability distribution, so its integral over all possible values of θ must equal 1. The proportionality constant is usually difficult or impossible to compute except in special cases. Monte Carlo simulation methods are often employed to compute marginal distributions or other low-dimensional summaries of $p(\theta|\mathbf{x})$ when analytic computation is difficult. Exact analytic results are easily obtained when $p(\mathbf{x}|\theta)$ and $p(\theta)$ are conjugate. A likelihood function and prior distribution are conjugate if

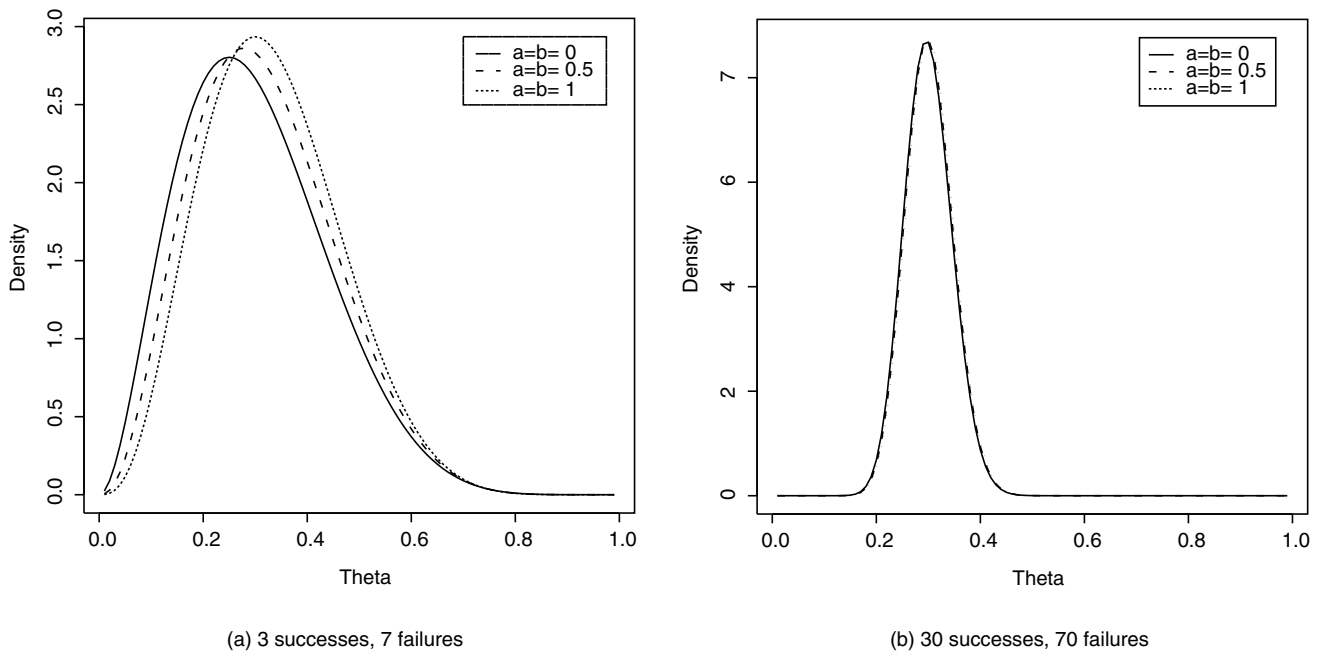


Figure 1 Posterior Distribution of θ Given Different Numbers of Successes and Failures and Different Beta Prior Distributions

they combine to form a posterior distribution from the same distributional family as the prior. The entry on PRIOR DISTRIBUTION lists several conjugate prior-likelihood pairs.

For an example, suppose one flips a coin n times, observes Y heads, and wishes to infer θ , the probability that the coin shows heads with each flip. The likelihood function is the binomial likelihood

$$p(Y|\theta) = \binom{n}{Y} \theta^Y (1 - \theta)^{n-Y}.$$

If one assumes the beta (a, b) prior distribution,

$$p(\theta) = \frac{\theta^{a-1}(1 - \theta)^{b-1}}{\beta(a, b)}, \quad \theta \in (0, 1),$$

which is conjugate to the binomial likelihood, then the posterior distribution

$$\begin{aligned} p(\theta | Y) &\propto \binom{n}{Y} \theta^Y (1 - \theta)^{n-Y} \frac{\theta^{a-1}(1 - \theta)^{b-1}}{\beta(a, b)} \\ &\propto \theta^{Y+a-1} (1 - \theta)^{n-Y+b-1} \end{aligned}$$

is proportional to the beta $(a + Y, b + n - Y)$ distribution. Conjugate priors eliminate the need to perform complicated integrations because the family of the

posterior distribution is known. One merely needs to examine the unnormalized distribution to determine its parameters.

The parameters of the beta distribution are interpretable as counts of successes and failures. Bayesian updating simply adds the number of successes to a and the number of failures to b . Interpreting parameters as prior observations provides a means of understanding the sensitivity of the posterior distribution to the prior. The beta (a, b) prior adds $a + b$ observations worth of information to the data. A Bayesian analysis requires that one specify a and b , but if there are a large number of successes and failures in the data set relative to a and b , then the choice of prior parameters will have little impact on the posterior distribution. In small samples, the choice of prior parameters can noticeably affect the posterior distribution.

Figure 1 plots the posterior distributions of θ given different numbers of successes and failures and different prior distributions. These distributions can be used to construct point estimates of θ using the posterior mean, median, or mode. Interval estimates of θ (known as credible intervals) may be obtained by locating the $\alpha/2$ and $1 - \alpha/2$ quantiles of the relevant beta distribution or by the method of highest posterior density (HPD). The posterior distribution for θ in Figure 1(a) exhibits desirable qualities, despite its sensitivity to

the prior. It is unimodal, constrained to lie within the interval (0, 1), and is skewed to reflect the asymmetry in the number of successes and failures. Prior sensitivity is usually a signal that the data provide only limited information and that frequentist methods will also encounter difficulty. For example, the standard frequentist method for Figure 1(a) centers on the point estimate $\hat{\theta} = .3$ with a standard error of roughly .15. If $\alpha < .05$, then the frequentist approximation produces confidence intervals that include negative values.

For more on the use of posterior distributions in Bayesian inference, consult Gelman, Carlin, Stern, and Rubin (1995) or Carlin and Louis (1998).

—Steven L. Scott

REFERENCES

- Carlin, B. P., & Louis, T. A. (1998). *Bayes and empirical Bayes methods for data analysis*. London: Chapman & Hall/CRC.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

POSTMODERN ETHNOGRAPHY

Postmodern ethnography is an ETHNOGRAPHY that goes beyond its traditional counterpart by expanding its REFLEXIVITY and its commitment to the groups it studies; it questions its truth claims and experiments with new modes of reporting. Postmodern ethnography derives its new goals from POSTMODERNISM. Postmodernism comprises a set of beliefs that span many disciplines, from architecture to literary criticism, from anthropology to sociology. It rejects grand narratives and meta-theories of modernism, with their preconceived and sweeping statements. Postmodernism proposes to “fragment” these narratives and study each fragment on its own. It does so, in sociology, by studying the interactions and details of everyday life and by analyzing each one separately, based on its situation and context.

Reflexivity is a paramount concern of postmodern ethnography, based on the belief that for too long, the role and power of the researcher have been ignored by ethnographers. The researcher chooses what to observe, when to observe it, what is relevant, and what is worth reporting. Whenever quotes from the members of society are used, they are snippets tailored

to support the argument put forth by the researcher. Yet, the study ignores these problems and tends to claim neutrality and objectivity, under the guise of letting the “natives” (subjects being studied) speak for themselves. Postmodern ethnography attempts to remedy these problems by reporting about the role and decision making of the researcher. Thus, although the ethnographer still controls the selectivity process, at least the readers are informed on how it is carried out. Furthermore, postmodern ethnography attempts to allow the voices of the “natives” to be heard to a greater degree. Various reporting modes, discussed below, are employed toward this end.

Commitment is another important element of postmodern ethnography. Traditional ethnographers tend to study the weak and the oppressed, yet the information obtained is rarely used to ameliorate the plight of the group studied. Postmodern ethnographers openly pledge political commitment to those studied and promise to share their findings to somehow help those being studied. Thus, the researcher, rather than control the study and exploit the subject for his or her own purpose, becomes a partner with those being studied, in a quest for their betterment.

A third concern of postmodern ethnography focuses on truth claims. Traditional ethnographies tend to be written in the “I was there” style (Geertz, 1988; Van Maanen, 1988). That is, the legitimacy and credibility of the study are achieved through the “eyewitnessing” of the researcher. He (or, seldom, she) narrates the ethnography by summarizing and categorizing the concerns of the subjects. The ethnographer inserts appropriate illustrative quotations from field notes here and there to underscore his or her thesis and create the illusion that the “natives” are speaking for themselves. Postmodern ethnography questions the fact that closer proximity to the respondents makes one closer to the “truth.” In fact, it questions the very notion of an absolute truth, holding, instead, that all we can achieve is a relative, temporal perspective that is contextually and historically bound. All we can hope to reach in ethnography is the understanding of a slice of life as we see it and as told to us by the respondents.

The belief in the relativity and bias of ethnography influences postmodern modes of reporting. Rather than relying on the researcher’s (*qua* writer) authority, the postmodern ethnographers try to minimize it as much as possible. Some postmodern ethnographies allow many of the subjects to voice their own concerns and keep the author’s comments to a minimum

(Krieger, 1983). Others expand the ethnographic data to include movies, television, ballads, folk tales, painting, dreams, and more. Some write AUTOETHNOGRAPHIES, looking at the past through the eyes of the present, by concentrating on the feelings and reactions of the researcher to the interaction. A few are trying out poetry as a form of ethnographic reporting. Plays and performances have become popular modes of reporting postmodern ethnography. In addition, short story format has become popular among postmodern ethnographers. All of these modes are ways to experiment in the attempt to capture a larger audience and to depict the events in a more immediate and compelling way.

A type of postmodern ethnography relies on the teaching of Michael Foucault. According to Foucault (1985), society uses its powers to control its members by creating for them formal sets of knowledge through which they apprehend reality; for example, an atlas is a way to understand and describe the sky, a dental chart provides names and organization for small enamel objects in our mouth, and so on. These ethnographers focus on the ways in which various groups create and control reality for its members.

Some feminists are pursuing postmodern ethnography. They reject traditional ethnographies as male dominated and criticize the work of traditional ethnographers by showing their male-centered concerns and biases (Clough, 1998). They advocate the use of females studying females to eliminate male dominance and allow women to express themselves freely. Also, they reject what they consider the exploitation of male researchers and promote support and cooperation for the women studied.

A final type of postmodern ethnography is VIRTUAL ETHNOGRAPHY. Some researchers are using the Internet to study various subcultures by becoming members of various chat rooms or joining Internet groups. This type of ethnography no longer relies on face-to-face communication but studies subjects whose existence can only be apprehended through the electronic media.

Finally, postmodern researchers address concerns of RELIABILITY and VALIDITY for their alternative ethnographies. Although there is no clear consensus, a very flexible set of principles based on a combination of aesthetic and political beliefs seems to emerge. Postmodern ethnographers pursue a self-admittedly utopian paradigm, based on the understanding that writing is not neutral but political and that it does not merely describe the world but changes it. Postmodern ethnography wishes to describe the world in

an engaging way; it desires to ameliorate the plight of the underprivileged groups studied, and all along it wants to remain as faithful as possible to the views of the social reality of the individuals studied.

—Andrea Fontana

REFERENCES

- Clough, T. P. (1998). *The end(s) of ethnography: From realism to social criticism*. New York: Peter Lang.
- Foucault, M. (1985). *Power/knowledge: Selected interviews and other writings: 1972–1977* (C. Gordon, Ed.). New York: Pantheon.
- Geertz, C. (1988). *Works and lives: The anthropologist as author*. Stanford, CA: Stanford University Press.
- Krieger, S. (1983). *The mirror's dance: Identity in a women's community*. Philadelphia: Temple University Press.
- Markham, A. N. (1998). *Life online: Researching real experience in virtual space*. Walnut Creek, CA: AltaMira.
- Van Maanen, J. (1988). *Tales of the field: On writing ethnography*. Chicago: University of Chicago Press.

POSTMODERNISM

Postmodernism is a term used in a variety of intellectual and cultural areas in rather different ways. The use of the same word creates a confusing and misleading impression of unity and coherence in developments across societal, academic, and cultural fields. There are three broad ways of talking about postmodernism: as an orientation in architecture and art, as a way of describing contemporary (Western or possibly global) society, and as a philosophy. This article mainly focuses on postmodernism in the third sense (i.e., as a philosophy or intellectual style within the social sciences), but a brief overview of other ways of using the term is also called for, especially because they are often viewed as linked. Postmodernism is impossible to define; it even runs against its spirit to try to do so. Many people do, however, use the term to refer to a style of thinking/cultural orientation emphasizing fragmentation and instability of meaning: an unpacking of knowledge, rationality, social institutions, and language.

Postmodernism began to be used in the 1970s as a term for a traditionally and locally inspired approach in architecture, in contrast to the ahistoric and superrational functionalism. Postmodernism also involves the mixing of architectural styles. In the arts,

postmodernism represents a partial continuation but also a partial break with modernism. Characterized by a rejection of the idea of representation and an emphasis on the challenge and the unfamiliar (e.g., Picasso), modernism in the arts actually has some affinity with postmodern philosophy. Nevertheless, the postmodern idea of “de-differentiation” between different spheres—such as high culture and low (mass) culture—and the break with elite ideas of a specific space and function of art make it possible to point at some overall features of postmodernism as an overall label that seems to work broadly in a diversity of fields:

the effacement of the boundary between art and everyday life; the collapse of the hierarchical distinction between high and mass/popular culture; a stylistic promiscuity favouring eclecticism and the mixing of codes; parody, pastiche, irony, playfulness and the celebration of the surface “depthlessness” of culture; the decline of the originality/genius of the artistic producer and the assumption that art can only be repetitious. (Featherstone, 1988, p. 203)

To the extent that there is such a trend in the arts—and some skeptics say that postmodernism is a favored idea among certain art critics rather than among larger groups of artists—that is indicative of wider societal developments, it fits into the idea of postmodernism representing a new society or a period in societal development.

Postmodernism started to be used in the social sciences in the late 1970s, drawing on French and, to some extent, Anglo-Saxon philosophers and social scientists (e.g., Derrida, Foucault, Baudrillard, Lyotard, Jameson, Harvey, and Bauman). Some of these rejected the term *postmodernism*; Foucault also resisted the idea of a “postmodern society or a period.” In particular, Jameson and Harvey, but also to some extent Lyotard, have been influential in forming ideas about a new period in societal development.

With references to a distinct societal period, we move postmodernism from culture to social science. Many authors then talk about *postmodernity* instead of postmodernism, and we will follow this custom here. There are various opinions about the meaning of this epoch. Some say that postmodernity signals postindustrialism as well as postcapitalism, but others suggest that postmodernity refers to cultural changes *within* capitalism. The periodization

idea typically indicates some notion of sequentiality—postmodernism (postmodernity) comes after modernism (modernity), but what this means is debated: Opinions vary from postmodernism emerging from the 1920s to the time after the Second World War and to the more recent time. An influential author is Jameson (1983), who refers to

a general feeling that at some point following World War II a new kind of society begun to emerge (variously described as postindustrial society, multinational capitalism, consumer society, media society and so forth). New types of consumption, planned obsolescence; an ever more rapid rhythm of fashion and styling changes; the penetration of advertising, television and the media generally to a hitherto unparalleled degree throughout society; the replacement of the old tension between city and country, center and province, by the suburb and by universal standardization; the growth of the great networks of superhighways and the arrival of automobile culture—these are some of the features which would seem to mark a radical break with that older prewar society. (pp. 124–125)

Postmodernity presupposes modernity, an orientation associated with the Enlightenment project. The most common point of departure in describing modernity—and thus postmodernity—is to focus on rationality seen as guiding principle as well as an attainable objective for the modern project. Rationality is embodied especially in the certainty and precision of science and knowledge, as well as the far-reaching control and manipulation of nature that this knowledge makes possible. Postmodernity then often refers to the idea of a society characterized by widely dispersed cultural orientations that doubt the possibilities and blessings of this rationalizing project, manifested in science, formal organizations (bureaucracies), and social engineering. “Postmodern society” thus goes against the claimed certainty, objectivity, and progress orientation of modernity.

There are different opinions of the relationship between modernity/postmodernity in terms of sequentiality/overlap. Whether the emergence of postmodernity means replacement of modernity or the emergence of a stream existing parallel with and in opposition to it is debated. For many commentators, postmodernity is better seen as a kind of reaction or

comment on modernism than a developmental phase coming after and replacing modernism. Arguably, principles and practices of modernity still seem to dominate a great deal of contemporary Western society.

Postmodernity is sometimes linked to postmodernism as a basic way of thinking about the world. It is then argued that societal developments call for sensitizing vocabularies and ideas that can say something interesting about recent developments and contemporary society. Conventional (modernist) social science and its language are viewed as outdated, incapable of representing subtle but pervasive societal developments. But opponents argue that discussions around postmodernity and postmodernism should be kept separated and confusions avoided: Matters of periodization and identification of social trends are empirical issues, whereas postmodernism is based on philosophical ideas and preferences for intellectual style uncoupled from how to describe contemporary society and social changes. One should be clear when addressing postmodernity and when being postmodernist, it is said.

Leaving postmodernity and the relationship between the period and the intellectual style and moving over to the latter, we enter frameworks and ideas in strong opposition to conventional ideas of social theory and methodology. *Postmodernism*, in this sense, strongly overlaps what is referred to as *poststructuralism*, and the two terms are frequently, when it comes to intellectual agenda and commitments, used as synonyms. There is still a tendency that postmodernism—also when separated from the postmodernity/period issues—refers to a societal/cultural context (e.g., Lyotard, 1984), whereas poststructuralism has a more strict interest in philosophy, particularly around language issues. The points of departures of the two “isms” also differ; postmodernism, in the case of Lyotard (1984), departs from changes in the conditions of knowledge production, whereas poststructuralists move away from the ideas of French structuralism (e.g., de Saussure). But increasingly in social science, postmodernism and poststructuralism are viewed as referring to the same agenda, but with *postmodernism* being the more popular term.

Very briefly, postmodernism can be described as “an assault on unity” (Power, 1990). For postmodernists, social science is a humble and subjective enterprise, characterized by tentativeness, fragmentation, and indeterminacy. Actually, many postmodernists would hesitate in labeling their work as social science. The social and the science part becomes downplayed, if

not rejected, by those taking postmodernism into its more extreme positions. For postmodernists, the social, as conventionally understood—systems, structures, agents, shared meanings, a dominant social order, and so forth—is replaced by discourse (language with strong constructing effects), images, and simulations in circulation.

The label of *postmodernism* in social science often summarizes the following five assumptions and foci (Alvesson & Deetz, 2000; Smart, 2000):

- *The centrality of discourse—textuality—where the constitutive power of language is emphasized and “natural” objects are viewed as discursively produced.* This means that there are no objects or phenomena outside or independent of language. Instead, language is an active force shaping our worlds. Different discourses thus mean different social realities.

- *Fragmented identities—emphasizing subjectivity as a process and the death of the individual, autonomous, meaning-creating subject.* The discursive production of the individual replaces the conventional “essentialistic” understanding of people. Against the conventional, “modernist” view of the human being as an integrated whole—the origin of intentions, motives, and meaning—postmodernists suggest a decentered, fragmented, and discourse-driven subject. There is, for example, no true meaning or experience associated with being a “woman”—instead, dominating discourses produce different versions of “woman” (e.g., mother, worker, voter, sexual being) with different effects on temporal subjectivities.

- *The critique of the idea of representation, where the indecidabilities of language take precedence over language as a mirror of reality and a means for the transport of meaning.* It is not possible to represent an objective reality out there. Pure descriptions or a separation of documentation and analysis become impossible. We cannot control language, but must struggle with—or simply accept—the locally context-dependent, metaphorical, and constitutive nature of language. Issues around texts and authorship then become crucial in social science (as well as in other fields of cultural work).

- *The loss of foundations and the power of grand narratives, where an emphasis on multiple voices and local politics is favored over theoretical frameworks and large-scale political projects.* Against the ambition to produce valid general truths guiding long-term, broad-scaled social and scientific projects, knowledge

and politics are seen as local and governed by short-term, performance-oriented criteria. Sometimes multiple voices and diversity are privileged.

- *The power-knowledge connection, where the impossibilities in separating power from knowledge are assumed and knowledge loses a sense of innocence and neutrality.* Power and knowledge are not the same, but they are intertwined and dependent on each other. Knowledge always has power effects; it produces a particular version of reality and—through being picked up and affecting how people relate to their worlds—shapes it.

Taking these five ideas seriously leads to a drastic reconceptualization of the meaning of social studies. Some people expect and hope that “the postmodern turn,” leading to a stress on “the social construction of social reality, fluid as opposite to fixed identities of the self, and the partiality of truth will simply overtake the modernist assumptions of an objective reality” (Lincoln & Guba, 2000, p. 178).

Postmodernist ideas are pushed more or less far and in somewhat different directions. A possible distinction is between skeptical and affirmative postmodernism (Rosenau, 1992). The skeptical version promotes a “negative” agenda, based on the idea of the impossibility of establishing any truth. Representation becomes a matter of imposing an arbitrary meaning on something. Research becomes a matter of deconstruction, the tearing apart of texts by showing the contradictions, repressed meanings, and thus the fragility behind a superficial level of robustness and validity. This approach strongly discourages empirical work, at least as conventionally and positively understood.

Affirmative postmodernism also questions the idea of truth and validity but has a more positive view of social research. Playfulness, irony, humor, eclecticism, and methodological pluralism are celebrated. So is local knowledge (including a preference for situated knowledge and a rejection of the search for abstract, universal truths). This version is not antithetical to empirical work and empirical claims, but issues around description, interpretation, and researcher authority become problematic. This has triggered profound methodological debates in many fields of social and behavioral science, especially in anthropology (Clifford & Marcus, 1986) and feminism (Nicholson, 1990).

Postmodernism can be seen as a key part in a broader trend in philosophy and social science, summarized

as the “linguistic turn.” This means a radical reconceptualization of language: It is seen *not* as a passive medium for the transport of meaning, a fairly unproblematic part of social life and institutions and carefully mastered by the competent social scientist. Instead, language (discourse) becomes a crucial constituting social force to be targeted by social researchers, who are themselves forced to struggle with basic problems around representation and authorship.

The implications for social research are substantial but, as said, may be taken more or less far and may inspire and rejuvenate rather than finish empirical social science (Alvesson, 2002). Interviews can, for example, be seen not as simple reports of reality or sites for the expression of the meanings and experiences of the interviewees but as local settings in which dominant discourses constitute subjects and their responses. Ethnographies are less viewed as the authoritative reports of other cultures based on carefully carried out fieldwork and more as fictional texts in which the authorship and the constructions of those being studied matter as much or more than the phenomena “out there” claimed to be studied. Rather than capturing meanings, finding patterns, and arriving at firm conclusions, postmodernistically inspired social research shows the indecisiveness of meaning, gives space for multiple voices, and opens up for alternative readings.

—Mats Alvesson

REFERENCES

- Alvesson, M. (2002). *Postmodernism and social research*. Buckingham, UK: Open University Press.
- Alvesson, M., & Deetz, S. (2000). *Doing critical management research*. London: Sage.
- Clifford, J., & Marcus, G. E. (Eds.). (1986). *Writing culture*. Berkeley: University of California Press.
- Featherstone, M. (1988). In pursuit of the postmodernism: An introduction. *Theory, Culture & Society*, 5, 195–215.
- Jameson, F. (1983). Postmodernism and consumer society. In H. Foster (Ed.), *Postmodern culture*. London: Pluto.
- Lincoln, Y., & Guba, E. (2000). Paradigmatic controversies, contradictions, and emerging confluences. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 163–188). Thousand Oaks, CA: Sage.
- Lyotard, J.-F. (1984). *The postmodern condition: A report on knowledge*. Minneapolis: University of Minnesota Press.
- Nicholson, L. (Ed.). (1990). *Feminism/postmodernism*. New York: Routledge Kegan Paul.
- Power, M. (1990). Modernism, postmodernism and organisation. In J. Hassard & D. Pym (Eds.), *The theory and*

philosophy of organisations. London: Routledge Kegan Paul.

Rosenau, P. M. (1992). *Post-modernism and the social sciences: Insights, inroads and intrusions*. Princeton, NJ: Princeton University Press.

Smart, B. (2000). Postmodern social theory. In B. Turner (Ed.), *The Blackwell companion to social theory* (pp. 447–480). Oxford, UK: Blackwell.

POSTSTRUCTURALISM

The post–World War II French intellectual climate was characterized not only by existentialism and PHENOMENOLOGY, Marxism and psychoanalysis, but also by the structuralist writings of de Saussure. STRUCTURALISM may be distinguished from other traditions through its contention that knowledge cannot be grounded in individuals or their historically given situations. Structuralism also rejects empiricism for its failure to distinguish between the appearance of a phenomenon and its underlying “reality.” It is a “science of relationships,” with the emphasis of holistic structuralism being on the interconnections or interdependencies between different social relations within society.

What has become known as poststructuralism rejects the universalist aspirations of structuralism in, for example, the work of Lévi-Strauss. However, as with many such terms, poststructuralism can work to alleviate researchers of the need for engagement, as much as to define a distinct tradition of social thought.

Under this school of thought, a diversity of thinkers has been included, with the German philosopher Friedrich Nietzsche providing one clear influence. Nietzsche espoused the doctrine of perspectivism: That is, there is no transcendental vantage point from which one may view truth, and the external world is interpreted according to different beliefs and ideas whose VALIDITY is equal to one another.

The poststructuralist movement—among whom have been included Michel Foucault, Jacques Derrida, Jacques Lacan, Jean-François Lyotard, Gilles Deleuze, Hélène Cixous, Luce Irigaray, and Julia Kristeva—was to subject received wisdom to critical scrutiny under the banners of anti-humanism and anti-essentialism, together with an absence of a belief in reason unfolding in history. A resulting critique of the unity of the subject has provoked a strong critical reaction. To understand this and its implications for social research, take the

works of two thinkers who have themselves engaged in debate: Jacques Derrida and Michel Foucault.

Derrida’s DECONSTRUCTIONIST project is driven by the desire to expose the “dream” of Western philosophy to find the transparent subject. For de Saussure, signs were divided into two parts: the signifier (the sound) and the signified (the idea or concepts to which the sound refers). The meaning of terms was then argued to be fixed through language as a self-referential system. In contrast, Derrida argues that when we read a sign, its meaning is not immediately clear to us. Not only do signs refer to what is absent, so too is meaning. Meaning is then left to flow along a chain of signifiers, with the result that it can never be fixed or readily apparent. An inability to find a fixed point is based on the idea of *différance*: That is, meaning will always elude us because of the necessity of continual clarification and definition. More language is thus required for this purpose, but language can never be the final arbiter. As a result, signs can only be studied in terms of other signs whose meanings continually evade us. This leaves Derrida constantly on the lookout for those who desire an unmediated truth by smuggling into their formulations a “transcendental signified.”

With this critique in mind, methodology can be opened up to alternative readings. Science may, for example, be read as a series of attempts at obfuscation, as opposed to liberation through illumination. Values and interests can be exposed where they are assumed not to exist to demonstrate that perspectivism, not universality, lies deep within its assumptions. RHETORIC buries its presuppositions away from the gaze of those who have not served its lengthy apprenticeships, and we cannot allude to language to solve this problem for the principle of “undecidability” will always obtain. The idea of questioning the metaphysics of presence affects social research practice in terms of, for example, its use of “consciousness” and “intentionality” as explanatory frameworks in the study of human relations. Instead, the study rests on how the social world is fabricated through what is known as “inter-textuality.” TEXTS are seen as social practices, and observations become acts of writing, not the reportings of an independent reality.

To take research beyond ideas of intentionality without lapsing into structuralism takes us into the realms of what has been characterized as Foucault’s “interpretive analytics.” The basis on which representations are formed and the potential for a transparent relationship between language and reality are abandoned for

the study of discourses in which a politics of truth combines to constitute subject and object. To analyze discourses in this manner means insisting on the idea that the social world cannot be divided into two realms: the material and the mental. Discourses provide the limits to what may be experienced as well as the attribution of meanings to those experiences. Expressed in these terms, a social researcher would examine not only the forms of appropriation of discourses but also their limits of appropriation (the latter being a neglected feature of research influenced by Foucault).

As part of this overall strategy that saw Foucault's work move from archaeology to genealogy, the notion of "eventalization" sees the singularity of an event not in terms of some historical constant or anthropological invariant. Instead, the use of Foucault's methods provides for what is argued to be a subversive strategy that breaches the self-evident to demonstrate how "things" might have been otherwise. A series of possibilities, without any allusion to a unified subject or reason waiting to be discovered in history, may then emerge. The social researcher is thereby required to "rediscover" possibility rather than engage in transcendent critique: "the connections, encounters, supports, blockages, plays of forces, strategies and so on which at a given moment establish what subsequently counts as being self-evident, universal and necessary. In this sense one is indeed effecting a sort of multiplication or pluralization of causes" (Foucault, 1991, p. 76).

There is no doubt that poststructuralism provides a rich vein of thought that provides new insights into social life. At the same time, some of this work is not always amenable to assisting social researchers in their endeavors. That is not to say it is not of importance beyond such a relationship, simply that one must always seek in thought the possibilities for illumination, clarification, and engagement.

—Tim May

REFERENCES

- Foucault, M. (1991). Questions of method. In G. Burchell, C. Gordon, & P. Miller (Eds.), *The Foucault effect: Studies in governmentality* (pp. 73–86). London: Harvester Wheatsheaf.
- Game, A. (1991). *Undoing the social: Towards a deconstructive sociology*. Buckingham, UK: Open University Press.
- Kendall, G., & Wickham, G. (1999). *Using Foucault's methods*. London: Sage.

POWER OF A TEST

In the SIGNIFICANCE TESTING of a HYPOTHESIS, there are two kinds of mistakes an analyst may make: TYPE I ERROR and TYPE II ERROR. One minus the Type II error of a statistical test defines the power of the same test. The power of a test is related to the true POPULATION PARAMETER, the population VARIANCE of the distribution related to the test, the SAMPLE size, and the LEVEL OF SIGNIFICANCE for the test.

—Tim Futing Liao

POWER TRANSFORMATIONS. See POLYNOMIAL EQUATION

PRAGMATISM

Pragmatism, in the technical rather than everyday sense of the word, refers to a philosophical movement that emerged in the United States in the second half of the 19th century and continued to have an influence throughout the 20th century. The core idea of pragmatism is that the meaning of any concept is determined by its practical implications; and that the truth of any judgment is determined in and through practical activity, whether in the context of science or in life more generally.

These central ideas are usually connected with a few others. One is rejection of the assumption, shared by many RATIONALISTS and EMPIRICISTS, that we can find some indubitable foundation on which knowledge can be built. Thus, for Charles S. Peirce, generally recognized to be the founder of pragmatism, inquiry always starts from a problem that emerges out of previous experience, and always involves taking much of that experience for granted: Only what is open to *reasonable* doubt is questioned, not all that is open to *possible* doubt. Similarly, the knowledge produced by inquiry is always fallible; it cannot be known to be valid with absolute certainty. However, in the long run, inquiry is self-corrective: It is geared toward the discovery and rectification of errors.

Another central pragmatist idea is that all cognition must be viewed as part of the ongoing transaction

between individual organisms and their environments. This is at odds with the influential ancient Greek view that true knowledge is produced by spectators detached from everyday human life. Thus, for William James, another influential pragmatist philosopher, the function of cognition is not to copy reality but to satisfy the individual organism's needs and interests, enabling it to achieve a flourishing relationship with its environment. For him, truth consists of what it proves generally expedient to believe. Indeed, for pragmatists, even philosophy and the sciences have a pragmatic origin and function. Thus, John Dewey viewed the procedures of scientific inquiry as central to the democratic ideal and as the proper basis for societal development.

There is diversity as well as commonality within the pragmatist tradition. Where some have taken science as the model for rational thinking, others have adopted a more catholic approach. Where some have assumed a form of REALISM, others have insisted that we only ever have access to how things appear in experience, not to some reality beyond it.

During the 20th century, pragmatist ideas were developed by a number of influential philosophers, notably C. I. Lewis, W. van O. Quine, H. Putnam, R. Rorty, and S. Haack. Rorty's work, in particular, has generated much controversy, not least about whether it represents a genuine form of pragmatism or a break with the older tradition.

Pragmatism has had considerable influence on thinking about social research methodology, notably but not exclusively through the Chicago school of sociology. A valuable, early text that draws heavily on pragmatism is Kaplan's (1964) *The Conduct of Inquiry*.

—Martyn Hammersley

REFERENCES

- Haack, S. (1993). *Evidence and inquiry: Toward a reconstruction of epistemology*. Oxford, UK: Blackwell.
 Kaplan, A. (1964). *The conduct of inquiry*. New York: Chandler.
 Mounce, H. O. (1997). *The two pragmatisms: From Peirce to Rorty*. London: Routledge and Kegan Paul.

PRECODING

A *code* is a number (or letter or word) that will be recorded on a data file as the answer to a particular

Which of the following best describes your employment situation? (Please circle *one* answer):

1. Employed
2. Self-employed
3. In full-time education/training
4. Unemployed
5. Retired

Figure 1 A Precoded Variable on Employment Status

CLOSED-ENDED QUESTION from a particular respondent or informant in a quantitative study. An efficient study will code most of its information at the point of collection—precoding. The interviewer (or the informant, in a self-completion questionnaire) will circle a number corresponding to his or her response, and this number will be typed into the data file without further thought or coding. For example, Figure 1 shows a precoded “employment status” variable.

For precoding to be an effective procedure, two things are necessary:

- The categories must be unambiguous, with no overlap between them. The coding scheme in Figure 1 fails in this respect: It is possible to be both employed and in full-time education (on an army scholarship to university, for example) or full-time training (as an apprentice, for example).

- The categories must be exhaustive, covering the whole field. Again, Figure 1 fails in this respect: There is no category covering full-time homemakers. Some may put themselves down as “unemployed”; others may consider themselves “retired.” Those who consider homemaking and the care of children a job in its own right are constrained to these two answers, however, or to a misleading use of one of the “employed” categories. In other words, the survey question structures the social world in a way that may not be acceptable to some of the respondents—a not uncommon fault in precoded survey questions.

One way around this problem is to provide an “Other” category in which respondents write in their status if it is not reflected in any of the precoded

Please circle the number that best describes your <i>highest</i> educational qualification.	
GCE or equivalent	1
“O” level or equivalent	2
AS level or equivalent	3
A + level or equivalent	4
Higher education qualification lower than an honors degree	5
Honors degree or equivalent	6
Qualification higher than an honors degree	7

Figure 2 A Question on Educational Level

categories. The responses will then have to be “office coded” after the event—assigned to one of the existing categories, a new category, or a catchall “other” category. This takes additional time and effort, however, and its use is to be avoided whenever possible.

Mostly, we go for single-coded variables—sets of codes from which the respondent is required to select one and only one (see Figure 2). It is possible to handle multicoded variables—variables that carry more than one code (e.g., “Circle all the choices that apply to you from the following list”), but these need special thought at the data entry stage if they are to be interpretable in the analysis.

—Roger Sapsford

See also INTERVIEW SCHEDULE, QUESTIONNAIRE

PREDETERMINED VARIABLE

The exogenous variables and lagged ENDOGENOUS VARIABLES included in a SIMULTANEOUS EQUATION MODEL are typically called *predetermined variables*. The classification of variables is crucial for model IDENTIFICATION because there are endogenous variables among the independent (or explanatory) variables in a system of SIMULTANEOUS EQUATIONS.

To introduce the idea of predetermined variables, one must distinguish between endogenous variables and exogenous variables. *Endogenous variables* (or jointly determined variables) have outcomes values determined within the model. In other words, endogenous variables have values that are determined

through joint interaction with other variables within the model. One is interested in the behavior of endogenous variables because they are explained by the rest of the model. The term *endogeneity* (or *simultaneity*) reflects the simultaneous character and feedback mechanism of the model that determines these variables.

Exogenous variables differ from endogenous variables in that they contribute to provide explanations for the behavior of endogenous variables. These variables affect the outcome values of endogenous variables, but their own values are determined completely outside the model. In this sense, exogenous variables are CAUSAL, characterizing the model in which endogenous variables are determined, but values of exogenous variables are not RECIPROCALLY affected. In other words, no feedback interaction or simultaneity between exogenous variables is assumed.

Consider a very simple macroeconomic model that is specified as a simultaneous equation model with three equations:

$$C_t = \alpha_0 + \alpha_1 Y_t + \alpha_2 Y_{t-1} + \alpha_3 R_t + \varepsilon_{Ct}, \quad (1a)$$

$$I_t = \beta_0 + \beta_1 R_t + \varepsilon_{It}, \quad (1b)$$

$$Y_t = C_t + I_t + G_t. \quad (1c)$$

In this system of equations, contemporaneous consumption (C_t), investment (I_t), and income (Y_t) are endogenous because their values are explained by other variables and determined within the system. There is a feedback mechanism between the endogenous variables because C_t is conditional on Y_t in the behavioral equation (1a), and Y_t depends on C_t through the identity equation (1c) at the same time. Interest rates (R_t) and government spending (G_t) are exogenous variables because they explain the behavior of endogenous variables, but their values are determined by factors outside the system. We visualize no feedback mechanism between these two exogenous variables.

There is one LAGGED endogenous variable, Y_{t-1} , included in this simple macroeconomic model. The value of this lagged endogenous variable is determined by its past value. For the current value of income (Y_t), its past value (Y_{t-1}) is observed and predetermined. Thus, past endogenous variables may be placed in the same category as the exogenous variables. In other words, lagged endogenous variables and exogenous variables are two subsets of predetermined variables. More information on the classification of variables with respect to identifying and estimating a system of equations may be found in Kmenta (1986); Judge,

Griffiths, Hill, Lütkepohl, and Lee (1985); and Wooldridge (2003).

—Cheng-Lung Wang

REFERENCES

- Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., & Lee, T.-C. (1985). *The theory and practice of econometrics* (2nd ed.). New York: John Wiley.
- Kmenta, J. (1986). *Elements of econometrics* (2nd ed.). New York: Macmillan.
- Wooldridge, J. M. (2003). *Introductory econometrics: A modern approach* (2nd ed.). Cincinnati, OH: South-Western College Publishing.

PREDICTION

Prediction involves inference from known to unknown events, whether in the past, present, or future. In the natural sciences, theories often predict the existence of entities that are subsequently discovered. In the social sciences, prediction usually refers to statements about future outcomes. To be more than a lucky guess, a scientific prediction requires supporting empirical evidence and must be grounded in a body of theory. Thus, Randall Collins (1995) justifies his successful prediction of the collapse of communism by its grounding in geopolitical theory.

FORECASTING, generally used by economists in preference to *prediction*, is sometimes synonymous with it. Forecasting is also used in a more restricted sense to refer to predictions that hold true only for a concrete instance, in a given place and time frame: forecasting automobile sales in the United States over the next 5 years, for example. A forecast in this narrower sense is not generalizable beyond the specific case.

One form of prediction is *EXTRAPOLATION*, whereby past trends are projected into the future. To yield successful predictions, extrapolation assumes that the causes and social processes underlying the trends remain constant. Extrapolation is typically accurate only in the short run. Not all extrapolations, however, are intended to be predictive. They may be offered as possible scenarios, warning us to change our behavior lest the undesired outcome occur. Thus, extrapolations about the effects of global warming may be given not as predictions but as exhortations to reduce consumption of fossil fuels.

Retrodiction involves predicting backwards from outcomes to their causes, showing why an already known outcome was to be expected. Although retrodiction can be no more than the spurious wisdom of hindsight, there are good reasons why retrodiction can be valid even though prediction was difficult or impossible. Timur Kuran (1995) argues that a problem in predicting revolutions is *preference falsification*, that is, people misrepresenting their preferences because of perceived social pressures. The citizens of an oppressive regime may be reluctant to express any dissatisfaction to a social researcher. After the revolution has occurred, evidence of preexisting discontent becomes abundant. To improve our prediction of revolutions, we might therefore need to analyze sources of information other than standard social surveys (e.g., diaries and memoirs, confidential letters, and secret archives).

A key problem for social prediction is that predictions frequently interact with the phenomena they predict. In a *self-fulfilling prophecy*, the prediction comes true because it has been enunciated; conversely, a *self-negating prophecy's* enunciation ensures that it will be false. A practical solution is to insulate the prediction from the audiences that might react to it. In some cases, proclaiming self-fulfilling or self-negating prophecies may be a deliberate attempt to influence events so as to avert or mitigate adverse outcomes or increase the probability of desired outcomes.

Predicting future states of a system from its initial state requires that minor variations in initial states have a limited impact on future states. This does not hold true in *chaotic* systems (CHAOS THEORY), as in meteorology, in which undetectable differences in initial conditions produce widely divergent outcomes.

A major difficulty is the necessary unpredictability of future knowledge (Popper, 1957). We cannot rationally claim to have full predictive knowledge of that which has not yet been conceptualized or discovered.

—Alan Aldridge

REFERENCES

- Collins, R. (1995). Prediction in macrosociology: The case of the Soviet collapse. *American Journal of Sociology*, 100(6), 1552–1593.
- Kuran, T. (1995). The inevitability of future revolutionary surprises. *American Journal of Sociology*, 100(6), 1528–1551.
- Popper, K. R. (1957). *The poverty of historicism*. Boston: Beacon.

PREDICTION EQUATION

Any REGRESSION-like analysis provides two kinds of information. First, it tells us the estimated effect of one or more INDEPENDENT VARIABLES on a DEPENDENT VARIABLE. Estimates of COEFFICIENTS and STANDARD ERRORS suggest the relative strengths of effects and their estimation precision. Second, a regression-like analysis can provide an ability to predict values of a dependent variable. The *prediction equation* relates to this second kind of information. More specifically, the prediction equation is an equation that links the independent variables to predicted values of the dependent variable. In a LINEAR REGRESSION analysis, the prediction equation is formed by taking only the systematic components of the model, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$, and using these for calculating predictions of the dependent variable. In other forms of analysis, such as LOGIT, PROBIT, and other latent index models, the prediction equation is linked in NON-LINEAR fashion to a PROBABILITY function that enables prediction.

Generally, the average prediction error within a sample is a useful criterion for assessing model fit in a population. An effective prediction equation should produce predictions that, on average, lie close to the actual values of the sampled dependent variable. We can then test the significance of average prediction error in attempting to relate the model fit in the sample to a population.

A prediction equation can also be used as a FORECASTING tool for predicting *individual values* of a dependent variable that lie either within or outside of the sample. For example, suppose we have a linear regression for which we want to predict the value of $\hat{y}_i^0$, a single observation that lies either within or outside of the sample. Then we would use the prediction equation $\hat{y}_i^0 = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}^0 + \cdots + \hat{\beta}_k x_{ki}^0$. An effective prediction equation should be able to forecast accurately single observations that lie either inside or outside the sample. So forecasting individual observations is an alternative approach to assessing the quality of a model.

AVERAGE PREDICTION ERROR

In using a prediction equation for assessing *average* prediction error for a linear regression, there are two accepted standards: R^2 and the STANDARD ERROR OF

estimate (*SEE*, sometimes written as $\hat{\sigma}_e$). The R^2 statistic is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

Note that this calculation takes the sum of the squared prediction errors (calculated using the prediction equation) and then standardizes these prediction errors relative to a naive or worst-case prediction. Subtracting this quantity from 1 yields a measure of fit ranging from 0 to 1. At the 0 end of the scale, we can conclude that the model has no predictive ability. At the 1 end of the scale, we can conclude that the model predicts every observation without error. Values between 0 and 1 provide a scaled assessment of average predictive ability.

A disadvantage of R^2 for assessing average model prediction error is that it provides no indication of how much prediction error can be expected in the metric of the analysis. The R^2 statistic is scaled from 0 to 1, but for intermediate values, it is unclear how much actual prediction error will occur. For this reason, some analysts prefer using the STANDARD ERROR OF ESTIMATE (*SEE*). This is calculated as

$$SEE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{N - k}}.$$

Note that the *SEE* calculation takes the square root of the variance in prediction error, again calculated using the prediction equation. The result is an average prediction error scaled in the metric of the dependent variable. The value contained in the *SEE* statistic tells the analyst how much error can be expected, on average, in units of the dependent variable when using the prediction equation. This provides a more intuitive description of a model's predictive ability.

In relating average prediction error from the sample to the population using either R^2 or *SEE*, the analyst usually calculates an *F* statistic for the null hypothesis that the model has no predictive ability. In terms of R^2 , the null hypothesis is that the R^2 statistic equals zero. In terms of *SEE*, the null hypothesis is that the average prediction error is just the average deviation of observations about the mean. The two hypotheses are equivalent and enable the analyst to infer how much prediction error could be expected, on average, in repeated sampling from the same population.

INDIVIDUAL PREDICTION

Average prediction error, as gauged by R^2 and SEE , provide a measure of overall model fit. However, if the analyst is interested in using the prediction equation to forecast particular values of the dependent variable, then these statistics provide a poor assessment of how much prediction error can be expected. One reason is that observations nearer the ends of the distribution of the independent variables become increasingly atypical. Toward the ends of the distribution of X , we have less data and therefore more uncertainty about individual predictions. Individual prediction error gets larger the further an observation on X gets from the typical value of X . For example, suppose we are interested in predicting an individual observation of the dependent variable using the prediction equation $\hat{y}_i^0 = \hat{\beta}_0 + \hat{\beta}_1 x_i^0$. The prediction error for a single observation using this equation would be

$$SE[y_i^0 - \hat{y}_i^0] = SEE \sqrt{1 + \frac{1}{n} + \frac{(x_i^0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}.$$

Note that this statistic weights the average prediction error (SEE discussed above) by a function capturing the distance of the predicted value from the center of the distribution. A multivariate version of this statistic requires the use of linear algebra and is given in Greene (2003).

If we are interested in individual predictions for multiple observations, then the average error in individual predictions is used as a measure of forecasting accuracy. This is called the ROOT MEAN SQUARED error (RMSE) or sometimes Theil's U statistic. This statistic is given as

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i^0 - \hat{y}_i^0)^2},$$

where m is the number of individual predictions. Note that this is just the square root of the average sum of squared prediction errors. Smaller values of RMSE imply stronger forecasting capability when using a prediction equation.

—B. Dan Wood and Sung Ho Park

REFERENCES

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Mahwah, NJ: Prentice Hall.

King, G. (1990). Stochastic variation: A comment on Lewis-Beck and Skalaban's "The R -squared." *Political Analysis*, 2, 185–200.

Lewis-Beck, M. S., & Skalaban, A. (1990). The R -squared: Some straight talk. *Political Analysis*, 2, 153–171.

PREDICTOR VARIABLE

In every RESEARCH DESIGN, there is a set of INDEPENDENT VARIABLES that can be thought of as antecedent conditions that affect a DEPENDENT VARIABLE. These variables are either manipulated by the researcher (as in the case of EXPERIMENTS) or simply observed by the researcher (as in the case of SECONDARY ANALYSIS OF SURVEY DATA). If an antecedent condition is simply observed by the researcher (i.e., is not manipulated by them), it is often referred to as a predictor variable.

The distinction between predictor and independent variables is a function of the research design. In experimental studies, there can be a meaningful distinction between the two because the independent variable is presumed to be under the control of the researcher (i.e., it is the manipulation or treatment in the experiment), whereas the predictor variable is seen simply as an antecedent condition (e.g., the gender of a subject) that predicts values on the dependent variable. In non-experimental studies (e.g., CORRELATIONAL studies), however, the distinction is moot because the researcher manipulates none of the variables in the MODEL.

Consider the following two research designs in which the researcher is interested in assessing the effects of campaign advertisements on vote choice. In the first case, the researcher gathers data for a RANDOM SAMPLE of voters on each of the following variables: (a) whether the respondent was likely to vote for Candidate Smith, (b) gender, (c) age, (d) party identification, and (e) exposure to campaign advertisements for Candidate Smith. Using these data, the researcher then estimates the following REGRESSION equation:

$$\text{Vote for Smith} = \beta_0 + \beta_1 \text{Gender} + \beta_2 \text{Age} + \beta_3 \text{PID} + \beta_4 \text{Ads} + \varepsilon. \quad (1)$$

In the second case, the researcher gathers data for a random sample of voters on only (a) gender, (b) age, and (c) party identification. Voters are then randomly assigned to either a control or an experimental group. Voters in the experimental group are shown Candidate Smith's campaign advertisements, and those in the

control group view only product advertisements. After the treatment (viewing campaign advertisements or not) is administered, respondents in both groups are asked whether they would likely vote for Candidate Smith. On the basis of the results, the researcher estimates the regression equation given in (1).

Conventionally, *all* of the variables on the right-hand side of the regression equation in the first design are referred to as independent variables (or REGRESSORS). In the second design, however, we can distinguish between independent variables and predictor variables. The independent variable in this case is whether the respondent saw the candidate's advertisements. It is the key EXPLANATORY VARIABLE of interest and the one that is manipulated by the researcher. The predictor variables in the model, then, are (a) age, (b) gender, and (c) party identification. Each subject in the study has some value on these variables prior to the researcher applying the treatment (thus they are correctly seen as antecedent conditions), and the values are thought to predict a respondent's score on the dependent variable (thus the nomenclature "predictor" variable).

Because the distinction between predictor variables and independent variables is of no consequence in the statistical analysis of the data (note that in the above example, the statistical technique employed did not vary from the first design to the second), it is often ignored in practice. However, the distinction is important for the design of a study (i.e., what variable is the treatment in experimental designs) and in the interpretation and explanation of the results.

—Charles H. Ellis

REFERENCES

- Frankfort-Nachmias, C., & Nachmias, D. (1996). *Research methods in the social sciences* (5th ed.). New York: St. Martin's.
- Howell, D. C. (1999). *Fundamental statistics for the behavioral sciences* (4th ed.). Belmont, CA: Duxbury.
- Jaeger, R. M. (1993). *Statistics: A spectator sport* (2nd ed.). Newbury Park, CA: Sage.

PRETEST

A field pretest is a dress rehearsal for a SURVEY. These PILOT tests are extremely useful tools, allowing researchers to identify potential problems with survey

items and/or data collection PROTOCOLS prior to fielding a study. Although all surveys should undergo a trial run before they are launched, pretests are particularly valuable in preparation for large and/or complex surveys. They are imperative in any study that involves COMPUTER-ASSISTED DATA COLLECTION. Potentially costly mistakes can be identified and remedied during this phase of survey development.

Basically, a pretest involves collecting data from a relatively small number of RESPONDENTS using the data collection procedures specified for the study. Ideally, the pretest sample is selected from the study's SAMPLING FRAME. When this is not practical or feasible, respondents are often selected on the basis of convenience and availability. However, the more closely the pretest SAMPLE mirrors the survey sample in terms of the characteristics being studied, the better. For studies calling for INTERVIEWER administration, it is advisable to have at least three interviewers involved in pretesting the questionnaire.

Pretest responses are tabulated and the MARGINAL distribution of responses is examined for indications of potential problems—for example, little VARIATION in response (either for the VARIABLE as a whole or for subgroups of interest) or incorrect skip patterns. Items for which a large percentage of respondents either report that they do not know the answer to the question or refuse to respond should probably be revised and retested.

Although it is possible to pretest an instrument in all modes of administration—interviewer administered (by telephone and in person) and self-administered (by mail or electronically)—interviewer-mediated studies offer researchers opportunities to study the interaction between interviewer and respondent and identify questions that are not performing well. Once interviewers are briefed about the research goals of both the project in general and specific questions, they contact members of a small practice sample under conditions that mimic those of the larger study to follow.

The interviews are tape-recorded with the respondents' express permission. Later, specially trained CODING personnel listen to the tapes and "behavior CODE" the interaction between interviewer and respondent. For each question asked, both sides of the interaction are coded. On the interviewer side, it is noted whether the question was read exactly as worded, whether items that were not applicable to a particular respondent were navigated correctly, and whether PROBES or clarifications were offered. On

the respondent side, tallies are kept of how often respondents interrupted interviewers before the entire question was read or requested clarifications and definitions and whether the answers provided fit the response alternatives offered. BEHAVIOR CODING allows the identification of items that interviewers find difficult to administer in a standardized way, as well as questions or terminology that are not consistently understood by respondents. If any problems are found, the instrument or protocols can be revised prior to fielding the survey.

—Patricia M. Gallagher

REFERENCES

- Fowler, F. J., Jr. (1995). *Improving survey questions*. Thousand Oaks, CA: Sage.
- Oksenberg, L., Cannell, C., & Kalton, G. (1991). New strategies for pretesting survey questions. *Journal of Official Statistics*, 7(3), 349–365.

PRETEST SENSITIZATION

When participants in an EXPERIMENT are pretested, there is a risk that they will be alerted or sensitive to the variable that the pretest is measuring and that, as a result, their posttest scores will be affected. In other words, experimental subjects who are affected by pretest sensitization become sensitized to what the experiment is about, and this affects their behavior. The SOLOMON FOUR-GROUP DESIGN is sometimes used to deal with this problem.

—Alan Bryman

PRIMING

An individual's experiences in the environment temporarily activate concepts that are mentally represented. The activation of these concepts, which can include traits, schemata, attitudes, stereotypes, goals, moods, emotions, and behaviors, heightens their accessibility. These concepts are said to be *primed*; that is, they become more likely to influence one's subsequent thoughts, feelings, judgments, and behaviors. Priming also refers to an experimental technique that is used to simulate the activation of concepts that usually occurs through real-world experiences.

The term *priming* was first used by Karl Lashley (1951) to refer to the temporary internal activation of response tendencies. The process of priming begins when a mentally represented concept becomes activated through one's experiences in the environment. The level of activation does not immediately return to its baseline level; rather, it dissipates over time. This lingering activation can then affect subsequent responses. Thus, recently used concepts can be activated more quickly and easily than other concepts due to their heightened accessibility. Through this mechanism, priming creates a temporary state of internal readiness. Importantly, priming is a passive, cognitive process that does not require motivation or intent on the part of the perceiver.

Several types of priming techniques are commonly used in psychology research (see Bargh & Chartrand, 2000). Perhaps the simplest is *conceptual priming*. In this technique, the researcher activates one concept unobtrusively in a first task. The activated concept then automatically affects individuals' responses in a second, later task outside of their awareness or intent. A seminal experiment in social psychology using conceptual priming showed that activated concepts could affect future social judgments (Higgins, Rholes, & Jones, 1977). Participants were primed in a first task with positive or negative traits (e.g., reckless, adventurous). Those primed with positive traits formed more favorable impressions of a target person engaging in ambiguous behaviors than those primed with negative traits. There are two types of conceptual priming. In *supraliminal priming*, the individual is consciously aware of the priming stimuli but not aware of the priming stimuli's influence on his or her later responses. In *subliminal priming*, the individual is aware of neither the priming stimuli nor the influence those stimuli have on subsequent responses.

A second priming technique is *carryover priming*. In this type of priming, the researcher first has participants deliberately and consciously make use of the concept to be primed. This can include having participants think or write about a goal, mind-set, memory, person, or other topic that activates the to-be-primed construct. The individual then engages in a second, unrelated task. The construct that was primed in the first task influences the person's responses in the second task. Critically, he or she is not aware of this influence. Oftentimes, carryover priming is used when the key concept is too abstract or detailed to be conceptually primed.

A third priming technique is *sequential priming*. This method tests the associations between various mental representations. The researcher presents a prime stimulus, followed by a target stimulus. The participant must respond in some way to the target (e.g., pronouncing it, determining whether it is a word, evaluating it). To the extent that the prime facilitates the response to the target, there is an associative link in memory between the two constructs.

All priming research should end with a check for suspicion and awareness of the influence of the prime.

—Tanya L. Chartrand and Valerie E. Jefferis

REFERENCES

- Bargh, J. A., & Chartrand, T. L. (2000). Mind in the middle: A practical guide to priming and automaticity research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 253–285). New York: Cambridge University Press.
- Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, 13, 141–154.
- Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon symposium* (pp. 112–136). New York: John Wiley.

PRINCIPAL COMPONENTS ANALYSIS

Principal components analysis (PCA) is the workhorse of exploratory multivariate data analysis, especially in those cases when a researcher wants to gain an insight into and an overview of the relationships between a set of variables and evaluate individuals with respect to those variables. The basic technique is designed for continuous variables, but variants have been developed that cater for variables of categorical and ordinal MEASUREMENT LEVELS, as well as for sets of variables with mixtures of different types of measurement levels. In addition, the technique is used in conjunction with other techniques, such as REGRESSION ANALYSIS. In this entry, we will concentrate on standard PCA, but overviews of PCA at work in different contexts can, for instance, be found in the books by Joliffe (1986) and Jackson (1991), and an exposé of PCA for variables with different measurement levels is contained in Meulman and Heiser (2000).

THEORY

Suppose that we have the scores of I individuals on J variables and that the relationships between the variables are such that no variable can be perfectly predicted by all the remaining variables. Then these variables form the axes of a J -dimensional space, and the scores of the individuals on these J variables can be portrayed in this J -dimensional space. However, looking at a high-dimensional space is not easy; moreover, most of the variability of the high-dimensional arrangement of the individuals can often be displayed in a low-dimensional space without much loss in variability. As an example, we see in Figure 1 that the two-dimensional ellipse A of scores of Sample A can be reasonably well represented in one dimension by the first principal component, and one only needs to interpret the variability along this single dimension. However, for the scores of Sample B, the one-dimensional representation is much worse (i.e., the variance accounted for by the first principal component is much lower in Case B than in Case A, and interpreting a single dimension might not suffice in Case B).

The coordinate axes of the low-dimensional dimensional space are commonly called *components*. If the components are such that they successively account for most of the variability in the data, they are called *principal components*. The coordinates of the individuals on the components are called *component scores*. To interpret components, the coordinates for the variables

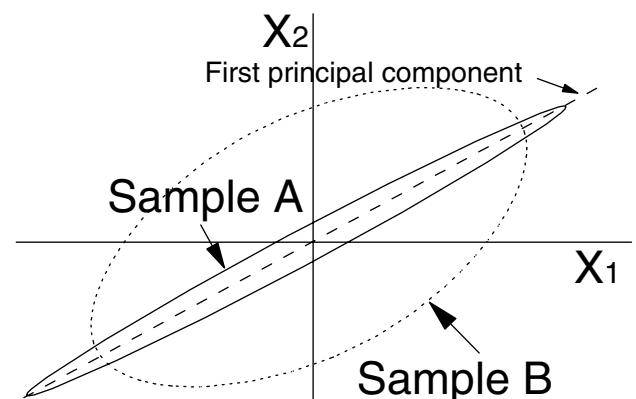


Figure 1 Two Samples, A and B, With Scores on Variables X_1 and X_2

NOTE: The ellipses indicate the approximate contour of the sample points. The first principal component is a good one-dimensional approximation to the scores of Sample A but much less so to the scores of Sample B.

on these components need to be derived as well, and the common approach to do this is via EIGENVALUE-EIGENVECTOR techniques. If both the variables and the components are standardized, the variable coordinates are the correlations between variables and components. By inspecting these correlations, commonly known as *component loadings*, one may assess the extent to which the components measure the same quantities as (groups of) variables. In particular, when a group of variables has high correlations with a component, the component has something in common with all of them, and on the basis of the substantive content of the variables, one may try to ascertain what the common element between the variables may be and hypothesize that the component is measuring this common element.

Unfortunately, the principal components do not always have high correlations with one specific group of variables because they are derived to account successively for the maximum variance of all the variables, not to generate optimal interpretability. This can also be seen in plots of the variables in the space determined by the components (see Figure 2). Such plots may show that there are other directions in this space for which components align with groups of variables and allow for a relatively simpler interpretation. Therefore, principal components are often rotated or transformed in such a way that the rotated components align with groups of variables if they exist. Such transformations do not affect the joint fit of the components to the data because they are restricted to the space of the principal components. When the rotation leads to correlated axes (oblique rotation; see the example), the coefficients reported for the variables are no longer correlations and are often called *pattern coefficients*. When the (rotated) components are uncorrelated, the square of each variable-component correlation is the size of the explained variance of a variable by a component, and the sum over all derived components of such squared correlations for a variable is the squared multiple correlation (often called COMMUNALITY) for predicting that variable by the derived components.

An essential part of any component analysis is the joint graphical portrayal of the relationships between the variables and the scores of the subjects through a BIPLLOT. For visual inspection, such plots should be at most three-dimensional, but when more dimensions are present, several plots need to be constructed (e.g., one plot for the first two dimensions and another for the third and fourth dimensions).

EXAMPLE

Suppose, as an extremely simplified example, that we have scores of nine children on an intelligence test consisting of four subtests and that a principal component analysis showed that two components are sufficient for an adequate description of the variability. Figure 2 (a biplot) gives the results of the two-dimensional solution, portraying the subtests A, B, C, and D as vectors and children 1 through 9 as points. Clearly, the subtests form two groups, and inspection of the contents of the subtests shows that A and D tap verbal intelligence and B and C numerical intelligence. The subtests do not align with the principal components, but an oblique transformation puts two axes right through the middle of the groups, and thus they may be hypothesized to represent numerical and verbal intelligence, respectively. The angle between the rotated axes indicates that they are correlated, which is not surprising considering all subtests measure intelligence. Due to their position near the positive side of the numerical intelligence axis, Children 6 and 8 did particularly well on the numeric subtests, whereas Children 3 and 5 scored considerably below average on the verbal subtests. Child 7 scores about average on all subtests, as she is close to the origin (see also the entry on BIPLOTS for rules of interpretation).

—Pieter M. Kroonenberg

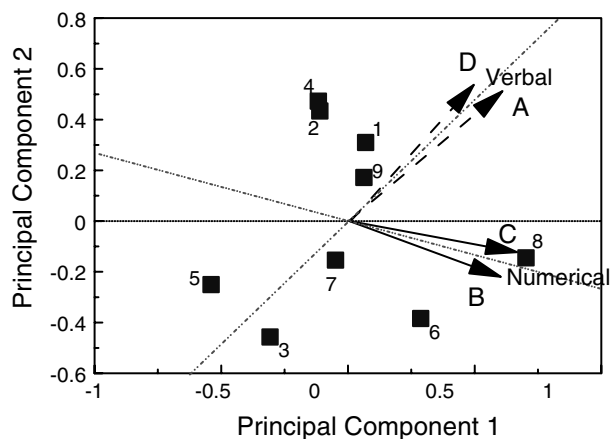


Figure 2 Example of a Biplot of the Four Subtests and Nine Children for the First Two Principal Components

NOTE: Subtests A and D measure numerical ability, and Subtests B and C measure verbal ability. The obliquely rotated components pass through each of the groups of subtests.

REFERENCES

- Jackson, J. E. (1991). *A user's guide to principal components*. New York: John Wiley.
- Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer.
- Meulman, J. J., & Heiser, W. J. (2000). *Categories*. Chicago: SPSS.

PRIOR DISTRIBUTION

A prior distribution $p(\theta)$ is a probability distribution describing one's subjective belief about an unknown quantity θ before observing a data set $\mathbf{x}$. BAYESIAN INFERENCE combines the prior distribution with the likelihood of the data to form the POSTERIOR DISTRIBUTION used to make inferences for θ . This entry describes the prior distribution's role in Bayesian inference from four points of view: subjective probability, objective Bayes, penalized likelihood, and hierarchical models.

At its foundation, Bayesian inference is a theory for updating one's subjective belief about θ upon observing $\mathbf{x}$. Subjectivists use probability distributions to formalize an individual's beliefs. Each individual's prior distribution is to be elicited by considering his or her willingness to engage in a series of hypothetical bets about the true value of θ . Models are required to make prior elicitation practical for continuous parameter spaces. Because of their computational convenience, *conjugate* priors are often used when they are available. A model has a conjugate prior if the prior and posterior distributions belong to the same family. Table 1 lists conjugate likelihood-prior relationships for several members of the exponential family. The prior parameters in many conjugate prior-likelihood families may be thought of as prior data, which provides a straightforward way to measure the strength of the prior (e.g., "one observation's worth of prior information"). See the entry on POSTERIOR DISTRIBUTION for an example calculation using conjugate priors.

One often finds that any reasonably weak prior has a negligible effect on the posterior distribution. Yet counterexamples exist, and critics of the Bayesian approach find great difficulty in prior elicitation (Efron, 1986). *Objective Bayesians* answer this criticism by deriving "reference priors," which attempt to model prior ignorance in standard situations. Kass and Wasserman (1996) review the extensive literature on

Table 1 Conjugate Prior Distributions for Several Common Likelihoods

Likelihood	Conjugate Prior
Univariate normal	Normal (mean parameter) gamma (inverse variance parameter)
Multivariate normal	Multivariate normal (mean vector) Wishart (inverse variance matrix)
Binomial	Beta
Poisson/exponential	Gamma
Multinomial	Dirichlet

reference priors. Many reference priors are improper (i.e., they integrate to ∞). For example, a common "noninformative" prior for mean parameters is the uniform prior $p(\theta) \propto 1$, which is obviously improper if the parameter space of θ is unbounded. Improper priors pose no difficulty so long as the likelihood is sufficiently well behaved for the posterior distribution to be proper, which usually happens in simple problems in which frequentist procedures perform adequately. However, when improper priors are used in complicated models, the propriety of the posterior distribution can be difficult to check (Hobert & Casella, 1996).

One difficulty with reference priors is that noninformative priors on one scale become informative after a change of variables. For example, a uniform prior on $\log \sigma^2$ becomes $p(\sigma^2) \propto 1/\sigma^2$ because of the Jacobian introduced by the log transformation. *Jeffreys' priors* are an important family of reference prior that are invariant to changes of variables. The general Jeffreys' prior for a model $p(\mathbf{x}|\theta)$ is $p(\theta) \propto \det(J(\theta))^{1/2}$, where $J(\theta)$ is the Fisher information from a single observation.

Although objective Bayesians strive to minimize the influence of the prior, practitioners of *penalized likelihood* actively seek out priors that will tame the highly variable parameter estimates that occur in high-dimensional models (Hastie, Tibshirani, & Friedman, 2001). For example, consider the regression model $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 I_n)$, where $\mathbf{X}$ is the $n \times p$ design matrix and $\mathbf{y}$ is the $n \times 1$ vector of responses. If $\mathbf{X}$ is ill-conditioned, then the least squares estimate of β is unreliable. An ill-conditioned design matrix can occur either because of an explicit collinearity in the data set or because $\mathbf{X}$ has been expanded by adding multiple interaction terms, polynomials, or other nonlinear basis functions such as splines.

A prior distribution $\beta \sim N(0, \tau^2 I_p)$ leads to a posterior distribution for β equivalent to ridge regression:

$$p(\beta | \mathbf{X}, \mathbf{y}, \tau, \sigma^2) = N(B, \Omega^{-1}).$$

The posterior precision (inverse variance) $\Omega = (\mathbf{X}^T \mathbf{X} / \sigma^2 + I_p / \tau^2)$ is the sum of the prior precision $\Omega_0 = I_p / \tau^2$ and the data precision $\Omega_1 = \mathbf{X}^T \mathbf{X} / \sigma^2$. If $\hat{\beta}$ is the least squares estimate of β , then the posterior mean B can be written as

$$B = \Omega^{-1} \Omega_1 \hat{\beta}.$$

Thus, B is a precision-weighted average of β and the prior mean of zero. It is a compromise between information in the prior and the likelihood. The prior distribution ensures that Ω is positive definite. When $\mathbf{X}$ is ill-conditioned, the prior stabilizes the matrix inversion of Ω , which dramatically reduces the posterior variance of β .

Finally, a prior distribution can be used as a mechanism to model relationships between complicated data structures. HIERARCHICAL (NON)LINEAR MODELS, which are often used to model nested data, provide a good example. A hierarchical model considers observations in a subgroup of the data (e.g., exam scores for students in the same school) to be conditionally independent given parameters for that subgroup. The parameters for the various subgroups are linked by a prior distribution whose parameters are known as hyperparameters. The prior allows subgroup parameters to “borrow strength” from other subgroups so that parameters of sparse subgroups can be accurately estimated.

To make the example concrete, let y_{ij} denote the score for student j in school i . Suppose exam scores for students in the same school are conditionally independent, given school parameters, with $y_{ij} \sim N(\theta_i, \sigma^2)$. A prior distribution that assumes each $\theta_i \sim N(\alpha, \tau^2)$ induces a posterior correlation on students in the same school. If one conditions on $(\alpha, \tau^2, \sigma^2)$ but averages over θ_i , then the correlation between any two individual scores in the same school is $\tau^2 / (\sigma^2 + \tau^2)$. Figure 1 provides a graphical description (known as the directed acyclic graph [DAG]) for the model.

To see how the prior allows one to “borrow strength” across population subsets, visualize information about (α, τ^2) flowing up from the data level. Then, given data and all other parameters, the posterior mean for θ_i is

$$E(\theta_i | y, \alpha, \sigma^2, \tau^2) = \frac{\frac{n_i}{\sigma^2} \bar{y}_i + \frac{1}{\tau^2} \alpha}{\frac{n_i}{\sigma^2} + \frac{1}{\tau^2}}.$$

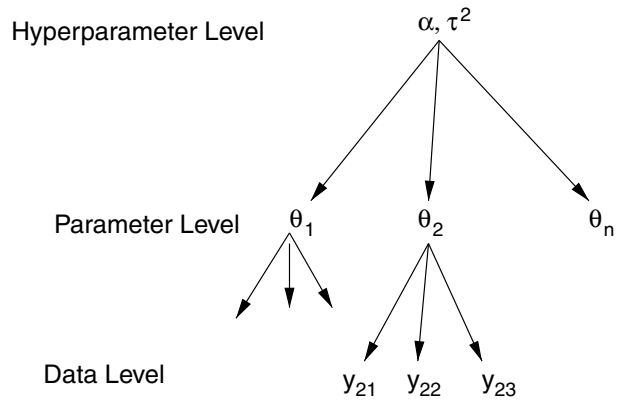


Figure 1 Graphical Depiction of a Hierarchical Model

NOTE: Each variable in the graph, given its parents, is conditionally independent of its nondescendants.

This expression is known as a *shrinkage estimate* because the data-based estimate is shrunk toward the prior. It combines information obtained from the entire sample, summarized by (α, τ^2) with data specific to subset i . The variation between subsets, measured by τ^2 , determines the amount of shrinkage. If $\tau^2 \rightarrow \infty$, then θ_i is estimated with the sample mean from subset i , but if $\tau^2 \rightarrow 0$, then all subsets share a common mean α . The posterior variance for θ_i is

$$\text{Var}(\theta_i | y, \alpha, \tau^2, \sigma^2) = \left(\frac{n_i}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1}.$$

In words, the posterior precision is the sum of the data precision and prior precision. When n_i is small, shrinkage estimation can greatly reduce the posterior variance of θ_i , compared to estimating each θ_i independently. For a complete discussion of hierarchical models, consult Carlin and Louis (1998).

—Steven L. Scott

REFERENCES

- Carlin, B. P., & Louis, T. A. (1998). *Bayes and empirical Bayes methods for data analysis*. London: Chapman & Hall/CRC.
- Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, 40, 1–5.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.
- Hobert, J. P., & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461–1473.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343–1370.

PRIOR PROBABILITY

A certain kind of PROBABILISTIC reasoning uses the observed data to update PROBABILITIES that existed before the data were collected, the so-called prior probabilities. This approach is called BAYESIAN INFERENCE and is in sharp contrast to the other existing approach, which is called *frequentist*. In the frequentist approach, the probabilities themselves are considered fixed but unknown quantities that can be observed through the RELATIVE FREQUENCIES. If the Bayesian approach is applied to a certain event, the researcher is assumed to have some knowledge regarding the probability of the event—the *prior probability*—and this is converted, using the BAYES THEOREM, into a posterior probability. The prior probability expresses the knowledge available before the actual observations become known, and the posterior probability is a combination of the prior probability with the information in the data. This updating procedure uses the BAYES FACTOR.

The prior probabilities are similar to ordinary probabilities, and they are called prior only if the goal is to update them using new data. One situation when this approach seems appropriate is when the prior probability comes from a census or an earlier large SURVEY and updating is based on a recent, smaller survey. Although census data, in most cases, are considered more reliable than survey data, they may become outdated, and a newer survey may reflect changes in the society that have taken place since the census was conducted. In such cases, a combination of the old and new pieces of information is achieved by considering the probability of a certain fact, computed from the census, as a prior probability and updating it using the recent survey information. A similar situation occurs when the prior information is taken from a large survey.

For example, suppose that the efficiency of a retraining program for unemployed people is investigated, in terms of the chance (probability) that someone who finishes the program will find employment within a certain period of time. When the program is launched, its efficiency is monitored in great detail, and the data are used to determine employment probabilities for

different groups of unemployed people. After the first year of operation, to cut costs, complete monitoring is stopped, and only small-scale surveys are conducted to provide figures for the efficiency of the program. In this case, the probabilities from the complete monitoring are more reliable than those computed from the later surveys. On the other hand, the efficiency of the program may have changed in the meantime for a number of reasons, and the survey results may reflect this change. Therefore, it is appropriate to consider the original probabilities as prior probabilities and to update them using the new data.

The simplest updating procedure uses the so-called BAYES THEOREM and may be applied to update the prior probabilities of certain events that cannot occur at the same time (e.g., different religious affiliations) and may all be compatible with a further event (e.g., agreeing with a certain statement about the separation of state and church). If the former events have known (prior) probabilities (e.g., taken from census data) and it is known what the conditional probability is for someone belonging to one of the religions to agree with the statement, then if agreement with the statement is observed for a respondent, the probabilities of the different religious affiliations may be updated to produce new probabilities for the respondent.

—Tamás Rudas

REFERENCE

Iversen, G. R. (1984). *Bayesian statistical inference* (Quantitative Applications in the Social Sciences, 07–043). Beverly Hills, CA: Sage.

PRISONER'S DILEMMA

A prisoner's dilemma (PD) is a game used to study motives and behavior when people interact seeking personal gain. In a typical PD game, two players each choose between two behavioral options: cooperation and defection. The payoff each player receives depends on not only what that player chooses but also what the other chooses. Neither knows the other's choice until both have chosen. Payoffs are structured so that whatever the other player does, choosing to defect always produces the highest payoff to the chooser; yet, if both players defect, each gets less than if both cooperate—hence the dilemma.

Table 1 Typical Payoffs in a Prisoner's Dilemma

		Choices of Player 2	
		C ₂	D ₂
Choices of Player 1	C ₁	3,3	0,5
	D ₁	5,0	1,1

SOURCE: Adapted from Rapoport and Chammah (1965).
NOTE: The first number in each cell is the payoff for Player 1; the second is for Player 2. C₁ and C₂ represent the cooperating choices by Players 1 and 2, respectively; D₁ and D₂ represent the defecting choices.

The name *prisoner's dilemma* comes from an imaginary situation in which two prisoners committed a serious crime together, but the authorities only have proof that they committed a minor crime. The authorities offer each prisoner a deal: inform on the other and get a lighter sentence. If only one prisoner informs, he will get a very light sentence and the other a very heavy sentence. If both inform, each will get a moderately heavy sentence. If neither informs, each will get a moderately light sentence.

In research, PD payoffs are usually positive rather than negative. Standard payoffs are presented in Table 1.

Iterated PD games have repeated trials, and both players learn at the end of each trial what the other did. In an iterated PD, players can use their own behavior strategically to influence the other's behavior. It has been suggested that social competition for nature's resources is an iterated PD, and the strategy producing the highest payoff over trials should be favored by natural selection. Accordingly, Axelrod and Hamilton (1981) invited game theorists in economics, sociology, political science, and mathematics to propose strategies, which were then pitted against one another in iterated PD games. The strategy that proved best of the 14 often quite elaborate strategies proposed was *tit for tat*. This strategy is quite simple; it involves cooperating on the first trial, then doing whatever the other player did on the preceding trial. The success of tit for tat has led some to argue that a tendency to reciprocate is part of our evolutionary heritage.

One-trial PD games are used to study social motives, not strategies. Game theory, which assumes that each person is motivated solely to maximize personal gain, predicts universal defection in a one-trial PD. Yet, when people are placed in a one-trial PD, about one third to one half cooperate (Poundstone,

1992). This finding has cast doubt on the assumption that maximization of personal gain is the sole motive underlying interpersonal behavior.

—C. Daniel Batson

REFERENCES

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211, 1390–1396.
Poundstone, W. (1992). *Prisoner's dilemma*. New York: Doubleday.
Rapoport, A., & Chammah, A. M. (1965). *Prisoner's dilemma*. Ann Arbor: University of Michigan Press.

PRIVACY. See ETHICAL CODES; ETHICAL PRINCIPLES

PRIVACY AND CONFIDENTIALITY

Jing was alone, elderly, lonely, and in failing health. She was delighted to be visited by a courteous Mandarin-speaking researcher who asked about Jing's health and housing. Jing served tea and cookies and poured out her health problems, until it occurred to her that the city's financial problems might be solved by reducing benefits of indigent elderly in failing health. To prevent this from happening to her, she explained that her (nonexistent) children take care of her health care needs and that she is free to live with them as required. However, the researcher was simply a doctoral student seeking to understand the needs of Jing's community.

What devices did Jing use to regulate the interviewer's access to her? What may have accounted for Jing's initial granting of access and later curtailment of access? What did Jing assume about confidentiality? What are the implications for informed consent?

Privacy and *confidentiality* have been variously defined. Here, *privacy* refers to persons and to having control over the access by others to oneself. *Confidentiality* refers to *data* about oneself and to agreements regarding who may have access to such information.

These definitions take into account the subjective nature of privacy. Behavior that one person gladly discloses to others, another person wants hidden from others, similarly for data *about* such behavior.

Valid and ethical research requires sensitivity to factors that affect what research participants consider as private. Failure to recognize and respect these factors may result in refusal to participate in research or to respond honestly, yet many sensitive or personal issues need to be researched. How do researchers learn to respectfully investigate such issues? Regulations of human research may appear to suggest that one size fits all and provide little definitional or procedural guidance. However, regulations direct researchers to find solutions to privacy issues. Depending on the research context, privacy-related dilemmas may be idiosyncratic and require special insights into the privacy interests of the specific research population and ways to respect those interests. The following is intended to guide the pursuit of such insight.

PRIVACY

The meaning of privacy inheres in the culture and personal circumstances of the individual, the nature of the research, and the social and political environment in which the research occurs. Laufer and Wolfe (1977) present a theory of personal privacy that recognizes the manifold cultural, developmental, and situational elements through which individuals orchestrate their privacy. It recognizes people's desire to control the time, place, and nature of the information they give to others and to control the kinds of information or experiences that are proffered to them. This theory embodies four dimensions of privacy, as follows.

1. The *self-ego dimension* refers to development of autonomy and personal dignity. Young children dislike being alone or separated from their parents. In research settings, they regard parents as a welcome presence, not an invasion of their privacy. By middle childhood, children seek time alone to establish a sense of self and nurture new ideas that become the basis of self-esteem, personal strength, and dignity; they seek privacy, sometimes manifested as secret hideouts and "Keep Out" signs on bedroom doors. Teenagers are intensely private, self-conscious, and self-absorbed as they forge an identity separate from their parents. They are embarrassed to be interviewed about personal matters in the presence of others, especially their

parents. Adults continue to need time alone and may become adroit at subtly protecting their privacy. (In the anecdote above, Jing shifted to polite lies when she decided that was in her best interests.)

2. The *environmental dimension* includes socio-physical, cultural, and life cycle dimensions. *Socio-physical* elements are physical or technological elements that offer privacy (e.g., private rooms, telephones). *Cultural* elements include norms for achieving privacy (e.g., cultural acceptability of white lies). *Life cycle* elements vary with age, occupation, available technology, and changing economic and sociocultural patterns.

3. The *interpersonal dimension* refers to how interaction and information are managed. Social settings and their physical characteristics provide options for managing social interaction; physical and social boundaries can be used to control people's access to one another.

4. The *control-choice dimension* grows out of the other three dimensions. Gradually, persons learn to use personal, cultural, and physical resources to control their privacy. Events that would threaten one's privacy early in the development of control/choice are later easy to control and pose no threat to privacy.

Using the dimensions described above to guide one's inquiry, discussion with networks of persons who are acquainted with the research population can provide useful information. Such networks might include, for example, relevant local researchers, educators, and outreach workers (e.g., social workers, farm agents, public health nurses, and other professionals). Focus groups and other forms of community consultation are useful ways to learn how a given community might perceive the research, what might be objectionable to them, how it can be made more acceptable, and how to recruit and inform prospective subjects (see, e.g., Melton, Levine, Koocher, Rosenthal, & Thompson, 1988). Talking with formal or informal gatekeepers and stakeholders, volunteering to work for relevant community outreach services, or using ethnographic methods can also provide useful insight into a population's privacy concerns. As the anecdote about Jing illustrates, it is in the best interests of the research program to understand and respect the privacy interests of subjects, as people have a host of techniques for covertly protecting their privacy and thereby confounding the goals of the researcher.

CONFIDENTIALITY

Confidentiality is an extension of the concept of privacy. It refers to identifiable data (e.g., videotape of subjects, nonanonymous survey data, data containing enough description to permit deductive identification of subjects) and to agreements about how access to those data is controlled. Research planning should consider needs for confidentiality and how access will be controlled. It should incorporate those plans into the methodology, institutional review board (IRB) protocol, informed consent, data management, training of research assistants, and any contractual agreements with subsequent users of the data. If feasible, problems of confidentiality should be avoided by gathering data in anonymous form, that is, so that the identity of the subject is never recorded. Identifiers include any information such as address, license number, or Social Security number, which can connect data with a specific person. When personal data are gathered, subjects may be particularly interested in knowing what steps will be taken to prevent possible disclosure of information about them.

However, promises of confidentiality do not necessarily contribute much to some subjects' willingness to participate in research, and extensive assurances of confidentiality accompanying nonsensitive research actually raise suspicion and diminish willingness to participate (Singer, von Thurn, & Miller, 1995). Many respondents, especially minorities, do not trust promises of confidentiality (Turner, 1982). Creative solutions to problems of subject distrust include consultation and collaboration with trusted community leaders in useful service connected with the research.

To whom or what does *confidentiality* pertain? Regulations of human research define *human subjects* as individually identifiable (not anonymous) living individuals. For many purposes, however, the researcher may also consider the privacy interests of entire *communities*, organizations, or even the deceased and their relatives (see Andrews & Nelkin, 1998). Issues of professional courtesy and maintaining the goodwill of gatekeepers to research sites may be as important to science and to individual investigators as protecting individual human subjects.

Issues of confidentiality have many ramifications as the following vignette illustrates:

In response to funding agency requirements, Professor Freeshare stated in his proposal that he would make his data available to other scientists as soon as his results are published. Thus, others could build on his data, perhaps linking them to related data from the same sample or reanalyzing his original data by different methods.

"Not so fast, Dr. Freeshare!" said his IRB. "There is a troubling disconnect in your description of your project. In your informed consent statement you promise confidentiality, but you must attach unique identifiers if other scientists are to link to your data. True, data sharing increases benefit and decreases cost of research, but that doesn't justify breach of confidentiality."

Dr. Freeshare revised his protocol to read as follows: Rigorous standards for protecting confidentiality will be employed to de-identify data that are to be shared, employing sophisticated methods developed by the U.S. Bureau of Census and other federal agencies that release data to academic scientists and the general public for further analysis. There are two kinds of data that might be shared, tabular data (tables of grouped data) and micro-data (data files on individual subjects, showing values for each of selected variables). There are also various methods of ensuring confidentiality: (a) de-identify the data before they are released to the public or to an archive (restricted data); (b) restrict who, when, or where the data may be used subsequently (restricted access); or (c) a combination of the two methods. It is anticipated that micro-data files will be prepared for release to other qualified scientists after the data have been carefully scrutinized, employing the Checklist on Disclosure Potential of Proposed Data Releases developed by the Confidentiality and Data Access Committee (CDAC). Restriction of data will be carried out using appropriate techniques described in the Federal Committee on Statistical Methodology (FCSM) "primer" (Chapter 2, FCSM, 1994). If qualified scientists present proposals for important research requiring use of unique identifiers, "restricted access" arrangements will be made to conduct the needed analyses at this university by statisticians who are familiar with the data and who are paid the marginal costs of organizing and conducting those analyses for the secondary user.

PROBLEMS

Preventing careless disclosure of information about subjects requires training of research personnel and planning of how raw data will be kept secure. As soon as possible, unique identifiers should be removed from the data and, if necessary, code numbers substituted; any key linking names to code numbers should be kept off-site in a secure place.

There are various possible threats to confidentiality apart from the researcher's own careless disclosure of subjects' identity. These include the following:

- Deductive identification based on demographic information contained in the data
- Recognition of unique characteristics of individual subjects (e.g., from videotapes, qualitative descriptions, case studies)
- Hacker break-in to computer data files containing unique identifiers
- Subpoena of data
- Legally mandated reporting by investigator (e.g., of child or elder abuse)
- Linkage of related data sets (from different sources or from longitudinal sequences)
- Audit of data by the research sponsor

Researchers are responsible for identifying possible threats to the confidentiality of data they plan to gather, employing appropriate means of ensuring confidentiality, and communicating to potential subjects how confidentiality has been ensured or whether there remain risks to confidentiality. Researchers must not promise confidentiality they cannot ensure.

SOLUTIONS

A major literature describes ways of ensuring confidentiality. The following summarizes main approaches to ensuring confidentiality. Each approach is accompanied by two sources of information: (a) a key bibliographic reference and (b) key words to employ with an online search engine such as Google. The Internet has become a rapidly evolving resource as new confidentiality problems and solutions emerge.

Broadened Categories and Error Inoculation. Deductive identification is possible when someone who knows distinctive identifying demographic facts about an individual (e.g., age, income, medical status, ZIP code) wishes to pry into other information on that individual. This can be prevented by broadening

data categories to obscure data from unique individuals (e.g., a billionaire can be included in a category containing many subjects labeled "annual income of over \$100,000"), or data can be "error inoculated" so that no individual case is necessarily correct. (Boruch & Cecil, 1979; confidentiality raw data)

Unique Characteristics of Individual Subjects or Settings in Case Descriptions and Videotapes. These can be disguised by various means, but elimination of all possible identifying characteristics may destroy the scientific value of the data. A second precaution is to assume that identities can be guessed and to avoid use of pejorative or unnecessarily personal language. (Johnson, 1982; confidentiality qualitative data)

Subpoena of Data. This is a possibility whenever data might be considered material to legal proceedings. In the United States, a subpoena can be blocked with a Certificate of Confidentiality obtained by the researcher from the National Institutes of Health prior to gathering data that might be subpoenaed. Obviously, data on criminal activities of subjects would call for a Certificate of Confidentiality. Less obvious are data that might be useful in future class-action suits, child custody disputes, or other civil legal actions; such threats to confidentiality are next to impossible to anticipate. A safer way to protect data from subpoena is to remove all unique identifiers as soon as possible. (National Institutes of Health, 2003; confidentiality, subpoena of data)

Linkage of Data to Unique Identifiers. Anonymity is the best practice when feasible. However, linkage with identifiers may be needed to permit, for example, follow-up testing for sampling validity or longitudinal research. Temporarily identified responses, use of brokers, aliases, or file-linkage systems can permit linkages without disclosure of subject identity. (Campbell, Boruch, Schwartz, & Steinberg, 1977; confidentiality file linkage)

Audits of Data. Audits of data by research sponsors require that the funder be able to recontact subjects to ensure that the data were actually gathered from the ostensible subject. To satisfy this requirement while protecting the privacy interests of subjects, the researcher (or the researcher's institution) may develop contractual agreements with such secondary users about how they will ensure confidentiality. Researchers should disclose these arrangements in the informed consent so that potential subjects understand what

researchers may do with the data subsequent to the project. (Boruch & Cecil, 1979; confidentiality data sharing)

Legally Mandated Reporting Requirements. Legally mandated reporting requirements (e.g., of child abuse) vary in the United States by state and county with respect to what constitutes abuse and who are mandated reporters. These requirements pertain mostly to helping professionals (therapists, physicians, teachers); many researchers happen to also be helping professionals. Moreover, in nine states—Florida, Indiana, Kentucky, Minnesota, Nebraska, New Hampshire, New Jersey, New Mexico, and North Carolina—anyone who has reason to suspect child maltreatment must report. Child abuse may include a wide range of activities—for example, neglect, physical abuse, psychological abuse, sexual abuse, and child pornography. Informed consent for research that may reveal reportable incidents should state that confidentiality could be promised except if the researcher believes that there is reasonable suspicion that a child has been abused. Researchers who have reasonable suspicion of child abuse should consult professionals who have a better understanding of reporting requirements (e.g., nurses, social workers, physicians) before proceeding to report the incident to local authorities. (Kalichman, 1999; reporting child abuse)

Laws concerning prevention and reporting of elder abuse are rapidly evolving. Most elder abuse is perpetrated by family members. It may include slapping, bruising, sexually molesting, or restraining; financial exploitation without their consent for someone else's benefit; and failure to provide goods or services necessary to avoid physical harm, mental anguish, or mental illness. (Bonnie & Wallace, 2002; reporting elder abuse)

Protection of Computer Files. Researchers routinely store electronic data on their institution's server. Such data are vulnerable to hackers, masqueraders, users of the local-area network (LAN), and Trojan horses. It is unwise to store data with unique identifiers where unauthorized users can possibly gain access. Understanding of these risks and relevant safeguards is evolving rapidly; the best current sources are found online (Skoudis, 2002; computer files hackers; confidentiality)

Online Research. Surveys, focus groups, personality studies, experiments, ethnography, and other

kinds of social and behavioral research are increasingly being conducted online, enabling researchers to reach far-flung subjects quickly, inexpensively, around the clock, and without a research staff. Accompanying problems and solutions to issues of confidentiality change rapidly as new technologies evolve. There are emerging technologies for protecting communication and personal identity and emerging cohorts of technology-sophisticated privacy activists. However, policies and methods of data protection cannot evolve fast enough to anticipate emerging problems. There are already digital communication networks on a global scale; hackers with a laptop computer and Internet technology could download any electronic data stored on any server anywhere in the world. (Nosek, Banaji, & Greenwald, 2002; Internet-based research)

IMPLICATIONS

There is an implicit understanding within social science that subjects' privacy will be respected and access to identified data appropriately limited. But what are subjects' privacy interests, and what limitations are appropriate? If there are special privacy interests, then special precautions (e.g., a Certificate of Confidentiality, error inoculation) may be needed and should be discussed with subjects and gatekeepers. When subjects are situated differently from the researcher, the problem is compounded by the difficulty of gaining insight into the subject's situation and privacy interests. For example, does the subject trust researcher promises of confidentiality? Are there "snoopers" who would engage in deductive identification of subjects or who would subpoena data? Clarity about such issues is elusive and cannot be settled definitively in regulatory language or in articles such as this. Such issues require that researchers know the culture and circumstances of those they study.

—Joan E. Sieber

REFERENCES

- Andrews, L., & Nelkin, D. (1998). Do the dead have interests? Policy issues of research after life. *American Journal of Law & Medicine*, 24, 261.
- Bonnie, R. J., & Wallace, R. B. (2002). *Elder mistreatment: Abuse, neglect, and exploitation in an aging America*. Washington, DC: Joseph Henry Press.
- Boruch, R. F., & Cecil, J. S. (1979). *Assuring the confidentiality of social research data*. Philadelphia: University of Pennsylvania Press.

- Campbell, D. T., Boruch, R. F., Schwartz, R. D., & Steinberg, J. (1977). Confidentiality-preserving modes of access to files and to interfile exchange for useful statistical analysis. *Evaluation Quarterly*, 1, 269–300.
- Confidentiality and Data Access Committee. (1999). *Checklist on disclosure potential of proposed data releases*. Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office. Retrieved from www.fcs.gov/committees/cdac/checklist.799.doc
- Federal Committee on Statistical Methodology. (1994). *Report on statistical disclosure limitation methodology* (Statistical Policy Working Paper #22). Prepared by the Subcommittee on Disclosure Limitation Methodology. Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office. Retrieved from www.fcs.gov/working-papers/wp22.html
- Johnson, C. (1982). Risks in the publication of fieldwork. In J. E. Sieber (Ed.), *The ethics of social research: Fieldwork, regulation and publication* (pp. 71–92). New York: Springer-Verlag.
- Kalichman, S. C. (1999). *Mandated reporting of suspected child abuse: Ethics, law and policy*. Washington, DC: American Psychological Association.
- Laufer, R. S., & Wolfe, M. (1977). Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of Social Issues*, 33, 44–87.
- Melton, G. B., Levine, R. J., Koocher, G. P., Rosenthal, R., & Thompson, W. (1988). Community consultation in socially sensitive research. *American Psychologist*, 43, 573–581.
- National Institutes of Health. (2003, July 22). *Certificates of confidentiality*. Retrieved from <http://grants1.nih.gov/grants/policy/coc/index.htm>
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-research: Ethics, security, design, and control in psychological research on the Internet. *Journal of Social Issues*, 58, 161–176.
- Singer, E., von Thurn, D., & Miller, E. (1995). Confidentiality assurances and response. *Public Opinion Quarterly*, 59, 66–77.
- Skoudis, E. (2002). *Counter Hack: A step-by-step guide to computer attacks and effective defenses*. Upper Saddle, NJ: Prentice Hall.
- Turner, A. C. (1982). What subjects of survey research believe about confidentiality. In J. E. Sieber (Ed.), *The ethics of social research: Surveys and experiments* (pp. 151–165). New York: Springer-Verlag.

PROBABILISTIC

A model or an argument is probabilistic if it is STOCHASTIC (i.e., incorporates uncertainty) and uses the concept of PROBABILITY. Probabilistic models and

arguments are often used in scientific inquiry when the researcher does not have sufficient information to apply a DETERMINISTIC approach and the available information is best formulated in terms of probabilities. A probabilistic model for a social phenomenon is a set of assumptions that describe the likely behavior of the entities involved, in terms of relationships, associations, causation, or other concepts that each have their specific probabilities. A probabilistic argument is often based on a probabilistic model and is not aimed at proving that a certain fact is true or will be the result of the discussed actions or developments. Rather, it assesses the probabilities of certain implications and outcomes.

The probabilities applied in a probabilistic model or argument may be results of subjective considerations, previous knowledge, or experience, and the results of the analysis of the model or of the argument are often used to assess whether the probabilities assumed seem realistic. In such cases, the probabilistic argument is applied iteratively.

In the social sciences, there are two major situations when probabilistic models and arguments are used. In one of these, the exact laws governing the relationships between certain variables are not precisely known. For example, in most societies, sons of fathers with higher levels of education have better chances of achieving a high level of education than sons of fathers with lower levels of education do. Based on educational mobility data, the conditional probabilities that a son whose father has a given level of education achieves a basic, medium, or high level of education may be estimated. These probabilities may be further modified depending on other factors that potentially influence educational attainment, such as the mother's educational level, ethnic or religious background of the family, the type of the settlement where the family lives, and so forth. By taking the effects of several factors into account, the probabilities are refined and may yield a good approximation of reality. This is a probabilistic model because not all factors are taken into account and the effects are not precisely known; rather, they are described using probabilities. The predictions from the model may be compared to reality to get better estimates of the probabilities used in the model and to find out if further factors need to be taken into account.

The other important situation when probabilistic arguments and models are used is when the precise factors and their effects are of secondary interest only

and the researcher is more interested in PREDICTING, based on data from a SAMPLE, the behavior of the entire POPULATION. A typical example of this situation is when data from opinion polls are used to predict election results. If, for example, 25% of the population would vote for a candidate, then with a sample size of 2,500, the pollster has 0.95 probability of finding a result between 23% and 27%. Based on this fact, if the poll results in, say, 26% popularity of the candidate, the pollster has (24%, 28%) as a 95% CONFIDENCE INTERVAL for the true value.

—Tamás Rudas

REFERENCE

Bennett, D. J. (1998). *Randomness*. Cambridge, MA: Harvard University Press.

PROBABILISTIC GUTTMAN SCALING

Probabilistic Guttman scaling is an extension to the classical GUTTMAN SCALING method, which is deterministic. In contrast, a probabilistic treatment is stochastic, and there exist numerous variations, the most prominent and earliest of which is a method proposed by Proctor (1970).

—Alan Bryman

See also GUTTMAN SCALING

REFERENCE

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika*, 35, 73–78.

PROBABILITY

The modern theory of probability is one of the most widely applied areas of mathematics. Whenever an empirical investigation involves uncertainty, due to SAMPLING, insufficient understanding of the actual procedure or of the laws governing the observed phenomena, or for other reasons, the concept of probability may be applied. In general, probability is considered a numerical representation of how likely is the occurrence of certain observations.

HISTORY

The concept of probability developed from the investigation of the properties of various games, such as rolling dice, but in addition to the very practical desire of understanding how to win, it also incorporates deep philosophical thought. First, questions concerning the chance of certain results in gambling were computed, and these led to working out the rules associated with probability, but the concept of probability itself was not defined. It took quite some time to realize that the existence of a PROBABILISTIC background or framework cannot be automatically assumed, and the question, “What is probability?” became a central one. In the late 19th century, it became widely accepted among philosophers and mathematicians that the correct approach to define the fundamentals of a precise theory (not only for probability but also for many other concepts underlying our present scientific thinking) is the axiomatic approach. The axiomatic approach does not aim at precisely defining what is probability; rather, it postulates its properties and derives the entire theory from these. It was Kolmogorov, who, in the early 20th century, developed an axiomatic theory of probability. Although the philosophical thinking about the nature of probability continues to this date, Kolmogorov’s theory has been able to incorporate the newer developments, and this theory led to widespread applications in statistics.

EXAMPLES AND AXIOMS

Usually, several constructions possess the properties postulated in a system of axioms and can serve as models for the concept, but the concept itself is not derived from, rather is only illustrated by, these CONSTRUCTS. For probability, a simple illustration is rolling a die with all outcomes being equally likely. Then, each outcome has probability equal to $1/6$, the probability of having a value less than 4 (i.e., 1, 2, or 3) is equal to $1/2$, and the probability of having an even outcome (i.e., 2, 4, or 6) is also $1/2$. The probability of having an outcome less than 4 and even is $1/6$, as this only happens for 2. There are, however, events that are possible but that cannot occur at the same time. For example, having a value less than 4 and greater than 5 is not possible. The event that cannot occur is called the impossible event and is denoted by 0.

Another example is the so-called geometric probability, for which a model is obtained by shooting at a

target in such a way that hitting any point of the target has the same probability. Here, hitting the right-hand half has $1/2$ probability, just like hitting the left-hand half or hitting the upper half. The probability that one hits the right-hand side and the upper side, at the same time, is $1/4$.

In a precise theory, the events associated with the observation of an EXPERIMENT possess probabilities. To define the properties of probability, one has to define certain operations on these events. The product of two events A and B occurs if both A and B occur and is denoted by AB . For example, one event may be, when rolling a die, having a value less than 4; another event may be having an even value. The product of these is having a value that is less than 4 and is even, that is, 2. Or, an event may be hitting the right-hand side of the target, another may be hitting its upper half, and the product is hitting the upper right-hand side quarter of the table. If two events cannot occur at the same time, their product is the impossible event. Another operation is the sum of two events, denoted as $A + B$. The sum occurs if at least one of the two original events occurs. The sum of having a value less than 4 and of having an even value is having anything from among 1, 2, 3, 4, 6. The sum of the right-hand side of the target and of its upper half is three quarters of the target, obtained by omitting the lower left-hand side quarter.

One can even think of more complex operations on the events. For example, consider the upper right-hand corner of the target and denote it by A_1 , then the one eighth of the target immediately below it (let this be A_2), then the one sixteenth of the target next to it (A_3), then the one thirty-second of the target (A_4), and so forth. One has here infinitely many events, and it is intuitively clear that they also have a sum—namely, the left-hand side half of the target. This is illustrated in Figure 1. The sum of these infinitely many events is denoted as $A_1 + A_2 + A_3 + A_4 + \dots$. Probability theory also applies to these more general operations.

Now we are ready to formulate the axioms of probability. The probability is a system of numbers (the probabilities) assigned to events associated with an experiment. The probabilities are assigned to the events in such a way that

1. The probability of an event is a value between 0 and 1:

$$0 \leq P(A) \leq 1.$$

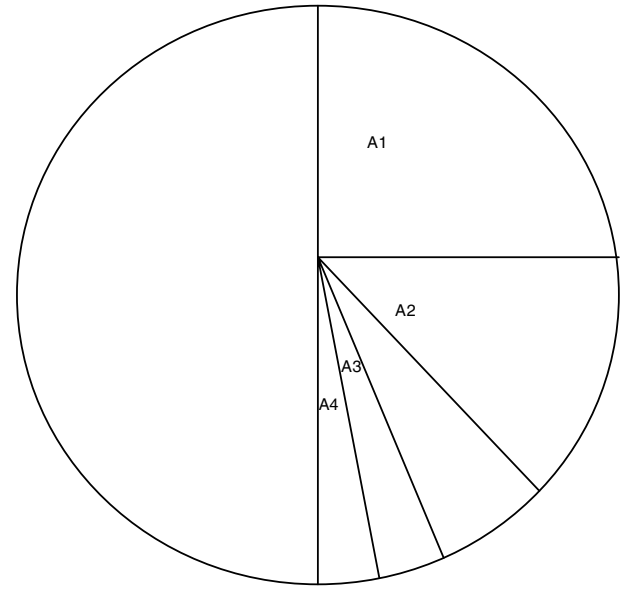


Figure 1 The Sum of Infinitely Many Events Illustrated on a Target

2. If an event occurs all the time (surely), its probability is 1:

$$P(S) = 1.$$

3. If two events are such that they cannot occur at the same time, then the probability of their sum is the sum of their probabilities:

$$P(A + B) = P(A) + P(B), \quad \text{if } AB = \emptyset.$$

The last property may be extended to cover the sum of infinitely many events, as in the case of hitting certain fractions of the target. A more general form of Axiom 3 postulates the following:

4. If a sequence of events $A_1, A_2, A_3, \dots$ is such that no two events can occur at the same time, then the probability of their sum is the sum of their probabilities:

$$\begin{aligned} P(A_1 + A_2 + A_3 + \dots) \\ = P(A_1) + P(A_2) + P(A_3) + \dots \\ \text{if } A_i A_j = \emptyset \text{ for } i \neq j. \end{aligned}$$

INDEPENDENCE

All properties of probability can be derived from Axioms 1 to 3 (or 1 – 3') and the further concepts

in probability theory that are precisely defined in this framework. One of the frequently used notions in probability theory is independence. The events A and B are independent if

$$P(AB) = P(A)P(B).$$

Notice that this formula is not the rule of computing the joint probability if independence is true; rather, it is the definition of independence. In the die-rolling example, the events of having an even number and of having a value less than 5 are independent because the former (2, 4, 6) has probability $1/2$ and the latter (1, 2, 3, 4) probability $2/3$, whereas the intersection (2, 4) has probability $1/3$, that is, their product. On the other hand, the events of having an even number and having something less than 4 are not independent because the former has probability $1/2$, the latter has probability $1/2$, and their product has probability $1/6$. The interpretation of this fact may be that within the first three numbers, even numbers are less likely than among the first six numbers. Indeed, among the first six numbers, three are even, but among the first three, only one is even.

CONDITIONAL PROBABILITY

The latter interpretation, in fact, points to another useful concept related to probability, called conditional probability. The conditional probability shows how the probability of an event changes if one knows that another event occurred. For example, when rolling the die, the probability of having an even number is $1/2$, because one half of all possible outcomes is even. If, however, one knows that the event of having a value less than 4 has occurred, then, knowing this, the probability of having an even number is different, and this new probability is called the conditional probability of having an even number, given that the number is less than 4. To see this, consider the list of possible outcomes of the experiment:

Possible outcomes: 1, 2, 3, 4, 5, 6
 Even numbers from among these: 2, 4, 6
 Probability of having an even number: $1/2$

Possible outcomes if the number is less than 4:
 1, 2, 3
 Even numbers from among these: 2
 Conditional probability of having an even number if the number is less than 4: $1/3$

The conditional probability of A given B is denoted by $P(A|B)$ and is precisely defined as

$$P(A|B) = P(AB)/P(B),$$

that is, the probability of the product of the events, divided by the probability of the condition. In the example above, we have seen that the probability of having a value that is even and also less than 4 is $1/6$, and the probability of having a value less than 4 is $1/2$, and their ratio is $1/3$.

It follows from the definition directly that if A and B are independent, then $P(A|B) = P(A)$; that is, conditioning on the occurrence of B does not change the probability of A if these are independent. Indeed, if A and B are independent,

$$\begin{aligned} P(A|B) &= P(AB)/P(B) \\ &= P(A)P(B)/P(B) = P(A), \end{aligned}$$

where the first equality is the definition of conditional probability, and the second one is based on the definition of independence. Therefore, independence of events means no relevance for each other. The probability of A is the same, whether or not B occurred, if A and B are independent.

In the target-shooting example, hitting the right-hand side half and hitting the lower half are easily seen to be independent of each other, by an argument similar to the one presented for the die-rolling example. This relies heavily on the fact that the two examples are similar in the sense that all outcomes are equally likely. In the case of rolling a die, the outcomes are the numbers 1, 2, 3, 4, 5, 6; in the target-shooting example, the outcomes are the points of the target. Although one is inclined to accept that the outcomes are equally likely in both cases, this assumption is straightforward for the six outcomes of die rolling but much less apparent for the infinitely many outcomes that are possible with the target shooting. It shows the strength of the underlying mathematical theory that such assumptions may be precisely incorporated into a probabilistic model. In the social sciences, models for experiments with infinitely many outcomes are used routinely; for example, approximations based on the NORMAL DISTRIBUTION assume that a variable may take on any number as its value, and there are infinitely many different numbers, just like infinitely many points on the target. In such cases, one uses a so-called density function to specify the probabilistic behavior.

APPLICATIONS

The foregoing simple examples are not meant to suggest that formulating and applying probabilistic models is always straightforward. Sometimes, even relatively simple models lead to technically complex analyses, but, more important, it is relatively easy to be mistaken on a basic conceptual level. The following example illustrates some of the dangers. The CEO of a corporation interviews three applicants for the position of secretary. The applicants—say, A, B, and C—know that three of them were interviewed and, as they have no information with respect to each other's qualifications, may assume that they all have equal chances of getting the job. Therefore, all of them think that the probability of being hired is $1/3$ for each of them. After the interviews were conducted, the CEO reached a decision, but this decision was not yet communicated to the applicants. Applicant A calls the CEO and says the following: As there are three interviewees and only one of them will be given the position, it is certain that from among B and C, there will be at least one who will not get the job. The CEO gives no information with respect to the chances of A by naming one from among B and C, who will not be given the position. The CEO tells A that C will certainly not get the position. After finishing the call, A thinks that now that C is out of the race, out of the remaining two candidates, A and B, each has $1/2$ probability of getting the job, that is, the probability changed from $1/3$ to $1/2$. Did or did not the CEO give information, with respect to the chances of A, by naming a person who would not get the job, if such a person exists, whether named or not? And what is the correct probability of A getting the job: $1/3$ or $1/2$? The situation is somewhat paradoxical, as one could argue for and also against both possible responses for both questions. The important point here is that assuming three candidates (and thinking that the probability is $1/3$) or assuming two candidates (and thinking that the probability is $1/2$) are two different probabilistic models and therefore cannot be directly compared. More precisely, asking whether $1/3$ or $1/2$ is the correct probability makes sense only if they are computed within the same model (and in that case, there is a clear answer). In the first model (three applicants), it is true that naming one who will not be given the job is not informative. But when the CEO named one candidate who would not be given the job, A defined, based on this information, a new probabilistic model, and in the new model (two applicants),

$1/2$ is the correct probability. The inference based on a probabilistic model (and this also applies to any kind of STATISTICAL INFERENCE) is only relevant if, in addition to the correct analysis of the model itself, it is also based on a model that appropriately describes the relevant aspects of reality. The results will largely depend on the choice of the model.

Probability theory has many applications in the social sciences. It is the basis of RANDOM SAMPLING, in which the units of the population of interest, usually people, are selected into the sample with probabilities specified in advance. The simplest case is SIMPLE RANDOM SAMPLING, in which every person has the same chance of being selected into the sample, and the steps of the selection process are independent from each other. In such cases, probability models with all outcomes having equal probability may be relevant, but the more complex sampling schemes used in practice require other models. Probabilistic methods are also used to model the effects of errors of MEASUREMENT and, more generally, to model the effects of not measured or unknown factors.

—Tamás Rudas

REFERENCES

- Bennett, D. J. (1998). *Randomness*. Cambridge, MA: Harvard University Press.
- Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed.). New York: John Wiley.
- Wonnacott, T. H., & Wonnacott, R. J. (1990). *Introductory statistics for business and economics* (4th ed.). San Francisco: Jossey-Bass.

PROBABILITY DENSITY FUNCTION

The probabilistic behavior of a CONTINUOUS VARIABLE is best described by its probability density function. Although in the case of a DISCRETE VARIABLE, the probabilities belonging to its different values can be specified separately, such a description of the behavior of a continuous variable is not possible. A continuous variable takes on all its values with zero PROBABILITY, and therefore specification of the individual probabilities would not be informative either. What is informative and, in fact, characterizes entirely the behavior of a continuous variable is the probability of having a value on an arbitrary interval. The

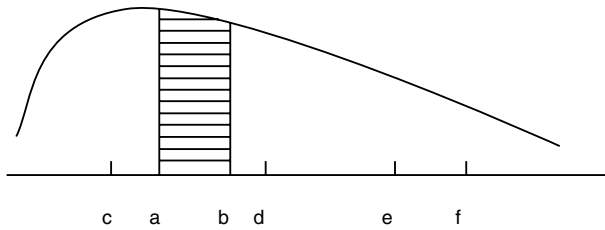


Figure 1 A Probability Density Function

probability of having an observation on an interval can be read off from the density function by determining the integral of the density function over this interval, that is, by taking the area under the function over this interval. This procedure is illustrated in Figure 1. The shaded area is the probability of observing a value on the interval (a, b) . The probability of other intervals is determined similarly. This probability depends on the length of the interval (the longer interval (c, d) has a higher probability) and on the location of the interval. For example, the interval (e, f) has the same length as (a, b) , but it is less likely to have an observation on it than on (a, b) . Interval probabilities determined by this method are easily seen to possess the properties of probability. For example, if an interval is the union of two other nonoverlapping intervals, then the area associated with it is the sum of the areas associated with the two intervals.

Density functions may also be used to determine the probability that a variable is smaller or greater than a certain value by calculating the relevant areas. To be a density function, the function, in addition to other properties, has to be nonnegative, and the total area under the function has to be equal to 1.

Density functions can be considered as being the theoretical counterparts of HISTOGRAMS, obtained by using shorter and shorter intervals. Parallel to this, the number of intervals increases and the boundary of the histogram becomes finer and finer.

The name of the density function comes from the fact that the probability of an interval determines how densely one can expect the observations on the interval. When the variable is observed, one can expect the number of observations on the interval to be proportional to the probability of the interval: On an interval with higher probability, the observations occur more densely than on an interval with lower probability.

The value of the density function for a possible value of the variable is called the *likelihood of the value*. For example, the likelihood of the value a is larger than the likelihood of the value e . In everyday language, the words *probability* and *likelihood* are used interchangeably, but in probability and statistics, they have different meanings. Both the values a and e have probability zero, but they have different likelihoods. A precise interpretation of the difference is that values in the vicinity of a are more likely to occur than values in the vicinity of e . It is also true that a short interval centered on a has a larger probability than an interval of the same length centered on e .

—Tamás Rudas

REFERENCE

Chatfield, C. (1983). *Statistics for technology* (3rd ed.). London: Chapman & Hall.

PROBABILITY SAMPLING. See RANDOM SAMPLING

PROBABILITY VALUE

The probability value (or p value) or the achieved probability level is a measure of the amount of evidence the data provide against a statistical HYPOTHESIS. It shows how extreme the data would be if the hypothesis were true. The p value is essentially the probability of observing data that are not more likely than the actually observed data set. Hypothesis tests are usually based on the p values. If the achieved probability level is smaller than a specified value—usually .05 but sometimes .01 or .10—the hypothesis is rejected. Most software packages routinely print the p values, and SIGNIFICANCE TESTING can be performed by comparing these values to the desired SIGNIFICANCE LEVELS.

As a simple example, consider the BINOMIAL TEST with $n = 10$ observations and the assumption that the success probability is $p = .5$. Out of the several properties of the observations, only the number of successes is relevant to test the hypothesis. Table 1 lists all possible results with their respective probabilities. If the data contain 7 successes (and 3 failures), the outcomes that

Table 1 Possible Results of the Binomial Test With $n = 10$ and $p = .5$

Number of Successes	Probability
0	0.000977
1	0.009766
2	0.04395
3	0.117188
4	0.205078
5	0.246094
6	0.205078
7	0.117188
8	0.04395
9	0.009766
10	0.000977

are more likely than the present outcome are those with 6, 5, or 4 successes. The outcomes that are as likely as or less likely than the actual data contain 0, 1, 2, 3, 7, 8, 9, or 10 successes; their combined probability is 0.3475, and this is the p value associated with the data and the test. Therefore, the actual data (with 7 successes) do not belong to the least likely 5% of the observations and do not suggest rejecting the hypothesis. If, on the other hand, the observations contain 1 success only, the outcomes that are as likely or less likely than this contain 0, 1, 9, or 10 successes, and their combined probability is 0.021484. This achieved probability level is less than .05, and the hypothesis is rejected because observing these data would be very unlikely if the hypothesis were true.

As another example, consider a 2×2 CONTINGENCY TABLE, and let the hypothesis be INDEPENDENCE. The data are given in Table 2. In this case, there are many possible independent tables, and the probability of a sample depends on which one of these is the true distribution. In such cases, MAXIMUM LIKELIHOOD ESTIMATION is used to select the distribution that would give the highest probability to the actual observations, and the p value is determined based on this distribution. To determine the probability value, we first compute

Table 2 Hypothetical Observations in a 2×2 Table

10	20
30	40

the value of a CHI-SQUARE statistic. The samples that have smaller likelihoods than the actual data are characterized by having a larger value of this STATISTIC. The combined probability of these samples (i.e., the p value) is equal to the probability that the statistic takes on a value greater than its actual value. The value of the Pearson CHI-SQUARE statistic is 0.793651 for the present data; the probability that the statistic exceeds this value is .37, and this is the p value of the data.

—Tamás Rudas

REFERENCE

Schervish, M. J. (1996). P -values: What they are and what are they not. *American Statistician*, 50, 203–206.

PROBING

Probing, a technique used during an in-person INTERVIEW to acquire additional information, is used to gather more explanation or clarification following an OPEN-ENDED QUESTION. It supplements the specific questions of the interview protocol. The technique allows the investigator to attain the precise information on which the research project is centered.

Research designs may employ different types of interview forms, varying in the degree of structure. At one end of the spectrum, there is a highly STRUCTURED INTERVIEW schedule, with the interviewer asking all respondents the same direct questions; protocols for these types of interviews may be very similar in design to a MAIL survey. A well-known example is the National Election Survey. At the other end, there is a much more loosely focused interview form, with the interviewer asking some general questions and allowing the interview to develop from there; such interviews have the feel of natural dialogue. More probing is required in less structured interviews. The interviewer will ask general questions and probe the respondent throughout the discussion to get at the information of interest.

Probes, as well as the primary interview questions, must not lead or suggest the interviewee's response. Probes should be simple prods for the interviewee to elaborate on his or her initial response. In interviewing, keep in mind that sometimes the most appropriate probe is a moment of silence; a pause alone is often enough to provoke the interviewee to develop the point.

Interviewees are different; some are much more loquacious than others, requiring little probing. Others

may be quite reticent. For example, if a respondent is asked if his or her organization used more than one institutional route to communicate its message and the interviewee responds with a simple “Oh we used every route in the book, you name it we tried it,” a probe could simply be, “Such as?” or “Could you describe that process a bit more?” prompting further explanation on the part of the interviewee.

—Christine Mahoney

REFERENCES

- Berry, J. (2002). Validity and reliability issues in elite interviewing. *PS—Political Science and Politics*, 35, 679–682.
- Dexter, L. A. (1970). *Elite and specialized interviewing*. Evanston, IL: Northwest University Press.

PROBIT ANALYSIS

Probit analysis originated as a method of analyzing quantal (dichotomous) responses. Quantal responses involve situations in which there is only one possible response to a stimulus, sometimes referred to as “all-or-nothing.” Early applications of probit analysis involved biological assays. A classic example involves the determination of death when an insect is exposed to an insecticide. Although theoretically the health of the insect deteriorates as exposure to the insecticide is increased, measurement limitations or focus on a specific result necessitate the use of a DICHOTOMOUS VARIABLE to represent the quantal response. Current applications in the social sciences often involve individual-level behavior such as whether a member of Congress votes for or against a bill or when a judge affirms or reverses a lower court decision.

Analytical work in the social sciences often uses a form of linear REGRESSION analysis (e.g., ORDINARY LEAST SQUARES [OLS]). A fundamental assumption of regression analysis is that the DEPENDENT VARIABLE is CONTINUOUS, which allows for a linear model. Many research questions, however, involve situations with a dichotomous dependent variable (such as the decision making of human actors, with a prime example being individual voting behavior). Because certain regression assumptions are violated when the dependent variable is dichotomous, which in turn may lead to serious errors of INFERENCE, analysis of dichotomous variables generally requires a nonlinear estimation procedure, and probit is a common choice.

The conceptual approach of regression is to estimate the coefficients for a linear model that minimizes the

sum of the squared errors between the observed and predicted values of the dependent variable given the values of the independent variables. In contrast, the approach of probit is to estimate the coefficients of a nonlinear model that maximizes the likelihood of observing the value of the dependent variable given the values of the independent variables. As such, probit estimates are obtained by the method known as MAXIMUM LIKELIHOOD ESTIMATION.

The probit model makes four basic assumptions. First, the error term is normally distributed. With this assumption, the probit estimates have a set of properties similar to those of OLS estimates (e.g., unbiasedness, normal sampling distributions, minimum variance). Second, the probability that the dependent variable takes on the value 1 (given the choices 0 and 1) is a function of the independent variables. Third, the observations of the dependent variable are statistically independent. This rules out SERIAL CORRELATION. Fourth, there are no perfect linear dependencies among the independent variables. As with OLS, near but not exact linear dependencies may result in MULTICOLLINEARITY.

Although probit estimates appear similar to those of OLS regression, their interpretation is less straightforward. Probit estimates are probabilities of observing a 0 or 1 in the dependent variable given the values of the independent variables for a particular observation. The coefficient estimate for each independent variable is the amount the probability of the dependent variable taking on the value 1 increases (or decreases if the estimate is negative), and it is given in standard deviations, a Z-SCORE, based on the cumulative normal probability distribution. Thus, the estimated probability that $y = 1$ for a given observation is

$$P(y = 1) = \Phi \left(\sum_{i=1}^n \hat{\beta}_0 + \hat{\beta}_i x_i \right),$$

where $\hat{\beta}_0$ is the estimated coefficient for the constant term, $\hat{\beta}_i$ are the estimated coefficients for the n independent variables (x_i), and Φ is the cumulative normal probability distribution.

Table 1 presents sample probit results for a model containing three independent variables, all of which are measured dichotomously (0/1). The probability that $y = 1$ for a particular observation is

$$\begin{aligned} P(y = 1) &= \Phi((2.31) * \text{constant} \\ &\quad + (-2.16) * x_1 + (-1.06) * x_2 + (-2.27) * x_3). \end{aligned}$$

Table 1 Sample Results of Probit Estimation

<i>Dependent Variable: y</i>		
<i>Independent Variables</i>	<i>Estimated Coefficient</i>	<i>Standard Error</i>
Constant	2.31	.56
x_1	-2.16	.42
x_2	-1.06	.45
x_3	-2.27	.80
N	107	
$-2LLR$ 3 df	65.28	

If all three of the x_i take on the value 1, the Z-score is -3.18. Computer programs may automatically calculate the probability, or one can look it up in a table to determine that the probability that $y = 1$ for this observation is only .0007. If all three of the x_i take on the value 0, the Z-score is 2.31, for a probability of .9896 that $y = 1$.

The marginal effect of a change in value for one of the x_i is often demonstrated by holding the other x_i at some value and calculating the change in probability based on the variable of interest. The values selected for the other x_i may be their means, although this can be theoretically troubling when the variables cannot actually take on such values, or some representative combination of values. For example, if we wish to know the effect of x_1 , then we might set $x_2 = 0$ and $x_3 = 1$ and calculate that if $x_1 = 1$, the probability that $y = 1$ is .0170, and the probability increases to .5160 when $x_1 = 0$. Note that the apparent effect of a particular variable will appear larger when the Z-score is closer to 0. For example, if the Z-score of the other variables equals 0, then a one standard deviation change will increase the probability from .5 to .8413. In contrast, if the starting Z-score is 2.0, then the same one standard deviation change will increase the probability from .9772 to .9987.

Significance tests of the estimated coefficients are achieved by dividing the estimate by its standard error to obtain a t -statistic. In the sample results for the first variable, $-2.16/.56 = -5.14$, which is significant at $p < .001$. The log-likelihood ratio is used to test the overall significance of the model. Specifically, $-2(\ln L_0 - \ln L_1)$, where $\ln L_1$ is the natural log of the likelihood value for the full model, and $\ln L_0$ is the value when all the variables except the constant are 0. The log-likelihood ratio, often denoted $-2LLR$, is a chi-square value that tests the hypothesis that all the coefficients except that of the constant are 0. The

degrees of freedom are the number of independent variables.

There are two basic types of GOODNESS-OF-FIT MEASURES for probit results. The first are the pseudo- R^2 measures that attempt to evaluate how much variance is explained by the model. One problem with such measures is that they do not consider the success of the model in correctly classifying the observations into the two categories of the dependent variable. This is particularly troublesome when there is a substantial imbalance in the percentage of observations in each category. Even a highly significant model will have difficulty correctly classifying observations into the nonmodal category when a large imbalance exists. An alternative measure considers how much the model reduces the error that would occur if one simply guessed the modal category for each observation. For example, if there were 70 observations in the modal category in a sample of 100, and the model correctly classified 80 observations, there would be a reduction of error of 33% (i.e., the model correctly classified 10 of the 30 errors one would make in simply guessing the modal category).

—Timothy M. Hagle

REFERENCES

- Aldrich, J. H., & Cnudde, C. F. (1975). Probing the bounds of conventional wisdom: A comparison of regression, probit, and discriminant analysis. *American Journal of Political Science*, 19, 571-608.
- Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit and probit models*. Beverly Hills, CA: Sage.
- Hagle, T. M., & Mitchell, G. E., II. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 36, 762-784.
- Liao, T. F. (1994). *Interpreting probability models: Logit, probit, and other generalized linear models*. Thousand Oaks, CA: Sage.
- Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.
- McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, 103-120.

PROJECTIVE TECHNIQUES

A projective technique is defined as an instrument that requires a subject to look at an ambiguous stimulus and to give it structure—for example, by saying what he

or she sees on an inkblot. It is assumed that the subject “projects” his or her personality into the response.

It is useful to contrast projective techniques with objective tests. The former have ambiguous stimuli, and the subject can, for example, see anything on an inkblot from an aadvark to a zebra. This great freedom of response enhances the projective nature of the response but can produce quantification problems. On an objective test, subjects might have to choose only between the answers “yes” and “no” in reference to self statements. Such limited responses enhance possibilities for statistical analysis. We might thus say that projective and objective tests are opposites in terms of their strengths and weaknesses. This has led some psychologists to suggest that they provide different kinds of information, with the projective test providing more about what is unique in the individual (IDIOGRAPHIC data) and the objective test telling us more about how the individual compares with a group in terms of personality traits (NOMOTHETIC data).

An example of the type of information provided by a projective technique might be the associations of the patient to having seen “four pink geese” on an inkblot. “When I was little I worked on my uncle’s farm—he would slaughter geese and the feathers would turn pink with blood. I recently remembered that he molested me. Funny—he had four daughters.” An example of the information provided by an objective test is finding that the score obtained on items measuring depression is very high when contrasted with the normative sample, suggesting that clinical depression is present.

Projective techniques have received much criticism from academic psychologists, usually pertaining to weak psychometric qualities. Nonetheless, they continue to be highly popular among clinically oriented psychologists.

Some of the better known projective techniques include the Rorschach technique, consisting of 10 inkblots in which the subject has to say what he or she sees. A second popular projective technique is the Thematic Apperception Test (TAT), in which the subject has to construct stories for ambiguous pictures of interpersonal scenes. Although there are scoring systems for the TAT and similar techniques, analysis is typically done impressionistically without using scores. A third common projective technique is the DAP (Draw-A-Person) test, in which the subject draws people and also possibly such things as a house, a tree, a family in action, and so forth. A fourth genre of projective techniques is an incomplete sentences test,

in which subjects have to complete a sentence stem, such as “most women. . . .” Answers to such items are usually analyzed impressionistically. (See SENTENCE COMPLETION TEST.)

Some recent trends in projective technique research include a standardization of different approaches to scoring and accompanying efforts to make the techniques more psychometrically sound (Exner, 2002), the development of alternate techniques designed for minority populations (Constantino, Malgady, & Vazquez, 1981), and use of the techniques as an actual part of the psychotherapy process (Aronow, Reznikoff, & Moreland, 1994).

—Edward Aronow

REFERENCES

- Aronow, E., Reznikoff, M., & Moreland, K. (1994). *The Rorschach technique*. Needham Heights, MA: Allyn & Bacon.
- Costantino, G., Malgady, R. G., & Vazquez, C. (1981). A comparison of the Murray-TAT and a new Thematic Apperception Test for urban Hispanic children. *Hispanic Journal of Behavior Sciences*, 3, 291–300.
- Exner, J. E. (2002). *The Rorschach: A comprehensive system* (Vol. 1). New York: John Wiley.

PROOF PROCEDURE

The proof procedure is a central feature of CONVERSATION ANALYSIS (CA), in which the analysis of what a turn at talk is doing is based on, or referred to, how it is responded to in the next speaker’s turn. Thus, a “request” may be identifiable as such, not simply on the basis of its content and grammar or on what an analyst may intuitively interpret it to be, but on the basis that it is treated as such in next turn. It may be constituted interactionally as a request, for example, by being complied with, or by noncompliance being accounted for, or by questioning its propriety, or by some other relevant orientation.

The proof procedure is not merely a method for checking the validity of an analysis that has already been produced. Rather, it is a basic principle of how CA proceeds in the first place. It is available as part of CA’s methodology because it is, in the first instance, a participants’ procedure, integral to how conversational participants themselves accomplish and check their own understandings of what they are saying and doing. Insofar as participants produce next turns on

the basis of what *they* take prior turns to be doing, analysts are able to use next turns for analysis (Sacks, Schegloff, & Jefferson, 1974). The proof procedure is therefore linked to the basic phenomena of conversational turn taking, including “conditional relevance” and “repair” (Schegloff, 1992); see entry on conversation analysis.

An important point to note here is that the proof procedure is *not* a way of establishing or proving what speakers may have “actually intended” to say. Next turns can be understood as displaying or implying participants’ hearings of prior turns, but such hearings may transform or subvert any such intentions rather than stand merely as hearers’ best efforts to understand them (cf. Schegloff, 1989, p. 200, on how “a next action can recast what has preceded”). But this is not a failure of the next-turn proof procedure because CA’s analytic goal is not speakers’ intentions, understood psychologically, but rather the methods (in the sense of ETHNOMETHODOLOGY) by which conversational interaction is made to work. Again, it is a participants’ option, having heard how a first turn has been responded to, to “repair” that displayed understanding in a subsequent “third turn.” What we then have is CA’s object of study—not speakers’ intentional meanings, but their interactional procedures for achieving orderly talk. Meaning is taken to be an interactional accomplishment, and the proof procedure takes us directly to the interactionally contingent character of meaning for participants, as well as the basis on which it can be studied in the empirically available details of recorded talk.

—Derek Edwards

REFERENCES

- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.
- Schegloff, E. A. (1989). Harvey Sacks—lectures 1964–1965: An introduction/memoir. *Human Studies*, 12, 185–209.
- Schegloff, E. A. (1992). Repair after next turn: The last structurally provided defence of intersubjectivity in conversation. *American Journal of Sociology*, 97(5), 1295–1345.

PROPENSITY SCORES

The propensity score is defined as the PROBABILITY of an individual being exposed to a new treatment, rather than the control treatment, conditional on a set of

covariates. Propensity scores are typically used in the context of CAUSAL inference, in which the objective is to estimate the effect of some intervention (the new treatment) by comparing the outcomes in two groups of subjects: one group that has been exposed to the intervention (the treated group) and an unexposed group (the control group).

Rosenbaum and Rubin (1983), who coined the term *propensity score*, show that subclassification, pair MATCHING, and model-based adjustment for the propensity score all lead to UNBIASED estimates of the treatment effect when treatment assignment is based on the set of COVARIATES that are in the propensity score. Subclassification involves classifying the observations (treated and control) into blocks based on the propensity score. Within each such block, the distribution of all covariates will be the same in the treated and control groups, at least in large SAMPLES. This outcome can be checked through statistical tests such as *T*-TESTS. Methods for causal inference in RANDOMIZED studies can then be used within each block. Pair matching chooses for each treated member the control member with the closest value of the propensity score. Analysis is done on the matched samples. The assumptions inherent in modeling (such as linearity) will be better satisfied in the matched groups than in the full treated and control samples because there is less EXTRAPOLATION. Other matching methods can also be used.

Conditional on the propensity score, treatment assignment is independent of the covariates, which implies that at the same value of the propensity score, the distribution of observed covariates is the same in the treated and control groups. This result justifies the use of the one-dimensional propensity score for matching or subclassification, in place of the multidimensional full set of covariates.

In randomized EXPERIMENTS, the propensity scores are known. In observational studies, however, they must be estimated from the data. Propensity scores are often estimated with LOGISTIC REGRESSION. Other classification methods such as CART or DISCRIMINANT ANALYSIS can also be used. Dehejia and Wahba (1999) provide an example of an observational study that uses propensity scores, including details on estimation.

OBSERVATIONAL studies designed by using subclassification and/or blocking on propensity scores resemble randomized experiments in two important ways. The first is that the analysis is done on treated and control groups with similar distributions of

the observed covariates. In randomized experiments, however, the distributions of all covariates are the same in the treated and control groups as a direct consequence of the randomization. Propensity scores can be used to replicate this as much as possible in observational studies.

The second way that an observational study designed using propensity scores resembles a randomized experiment is that the outcome data are not used in the design stage. The propensity score is estimated and assessed, and matched samples and/or subclasses are formed, without involving the outcome. Rubin (2001) discusses in more detail the use of propensity scores in the design of observational studies.

—Elizabeth A. Stuart

REFERENCES

- Dehejia, R. H., & Wahba, S. (1999). Causal effects in non-experimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053–1062.
- Rosenbaum, P. R. (2002). *Observational studies*. New York: Springer-Verlag.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2, 169–188.
- Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 659–706.

PROPORTIONAL REDUCTION OF ERROR (PRE)

Assume we have two models for predicting a variable Y and that for each MODEL, we have some measure of the inaccuracy of our PREDICTIONS. Also assume that Model 1 is the more “naive” or less informed model, for instance, predicting the value of Y without knowing anything about any of the predictors of Y ; Model 2 is a more informed model, using information about one or more predictors of Y to predict the value of Y . The measure of inaccuracy for Model 1, E_1 , may be a measure of the VARIATION of the observed values of Y about the predicted value of Y , an *actual count* of the number of ERRORS, or, when different

predictions under the same (naive) rule (e.g., given the number of cases in each category, randomly guessing which cases belonged in which category) would produce different numbers for E_1 , the *expected number* of errors based on the rule for prediction in Model 1. The measure of inaccuracy for Model 2, E_2 , would be the variation of the estimated value of Y in the second model or the *actual* number (not expected, because the more informed Model 2 should produce a unique prediction for each case) of errors of prediction made with the second model.

The *proportional change in error* is the ratio $(E_1 - E_2)/E_1$. As described in Menard (2002) and Wilson (1969), it is possible that the naive prediction of Model 1 may be better, by some measures of error, than the more informed prediction of Model 2, in which case we have a *proportional increase in error*. If the value of $(E_1 - E_2)/(E_1)$ is positive, however, it indicates that we are better able to predict the value of Y from Model 2 than from Model 1, and using Model 2 results in a *proportional reduction in error* (PRE) = $(E_1 - E_2)/(E_1)$. Multiplying the PRE by 100 gives us a *percentage reduction in error*, the percentage by which we reduce our error of prediction by using Model 2 instead of Model 1. The clear interpretation of PRE measures has led Costner (1965) and others to argue that PRE measures should be used in preference to other measures of association in social science research.

Examples of PRE measures include Goodman and Kruskal's τ_{YX} and λ_{YX} for bivariate association between nominal variables, in which Y is the variable being predicted and X is the predictor (Reynolds, 1984); the square of Kendall's τ_b , $(\tau_b)^2$, for bivariate ASSOCIATION between ORDINAL variables (Wilson, 1969); and the square of PEARSON'S r , r^2 , for ratio variables. For prediction tables, such as those generated by LOGISTIC REGRESSION or DISCRIMINANT ANALYSIS, PRE measures include λ_p , τ_p , and ϕ_p , similar but not identical to the corresponding measures for CONTINGENCY TABLES (Menard, 2002). For the special case of 2×2 contingency tables (but *not* more generally), several of these PRE measures produce the same numerical value: $r^2 = \tau_b^2 = \tau_{YX}^2$. In models with more than one predictor for a dependent variable, the familiar R^2 for ORDINARY LEAST SQUARES regression analysis and the LIKELIHOOD RATIO R^2 , R_L^2 for logistic regression analysis are also PRE measures (Menard, 2002).

—Scott Menard

REFERENCES

- Costner, H. L. (1965). Criteria for measures of association. *American Sociological Review*, 30, 341–353.
- Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Reynolds, H. T. (1984). *Analysis of nominal data* (2nd ed.). Beverly Hills, CA: Sage.
- Wilson, T. P. (1969). A proportional reduction in error interpretation for Kendall's tau-b. *Social Forces*, 47, 340–342.

PROTOCOL

Protocol has several usages in the literatures of social science methodology, but with all of them, the core meaning, generally construed, is “a detailed statement of procedures for all or part of a research or analytic undertaking.” This core meaning is found across fields ranging from psychiatry to sociology and from political science to epidemiology. (Note, however, that *protocol* also is a very common term in the literatures of computer and information science, and there its meaning is rather different—that is, “a set of rules for transferring information among diverse processing units.” The processing units in question will often be computers, but they may be sentient beings.)

One early social science use of the term was in research employing questionnaires (whether in broad sample surveys or in more focused interviewing), in which *interview protocol* was used synonymously with QUESTIONNAIRE or INTERVIEW SCHEDULE. Of course, here the detailed statement of procedures is limited entirely to matters of sequence and wording for the questions being asked.

Another early use was in ARTIFICIAL INTELLIGENCE models of problem solving, where “thinking-aloud protocols” or VERBAL PROTOCOLS were used first to provide human points of comparison with model behaviors and later to provide direct insight into how people reason (Ericsson & Simon, 1984; Newell & Simon, 1961, 1971). Thinking-aloud protocols represented participants’ step-by-step recounting of their analytic thoughts while in the process of solving puzzles or problems. These recountings became the raw data of an investigation. Perhaps because sophisticated (as opposed to, say, keyword) analysis of verbal protocols is such a daunting task, we now find little use of such protocols in most of the social sciences. However, they remain important in some areas of psychology (Carroll, 1997).

At the present time, an increasingly popular social science use of *protocol* is borrowed from medicine, where it has long been paired with TREATMENT or research. The meaning is akin to RESEARCH DESIGN, although a *treatment protocol* in medicine has infinitely more detail to it than is found in the typical non-experimental research design in the social sciences. Especially in clinical trials, a “protocol” covers every facet and every particular of all procedures in an effort to totally standardize data collection across times and sites. In contrast, the areas of standardization in research designs for the nonexperimental social sciences are for the most part focused loosely on measurement procedures. In many field studies, for example, there will be an effort to standardize questioning and the coding of answers. Fully standardized CODING will also be sought in the treatments of texts of one kind or another.

By their very nature, EXPERIMENTS require detailed research designs/protocols so that investigators can run multiple sessions under conditions that are as close as possible to true constancy. Hence, the growing use of experimental methods in many of the social sciences will be accompanied by increased use of explicit research protocols. Also, in an important contribution to the research enterprise, the publication of those protocols will make exact REPLICATIONS more common than they have been to date.

—Douglas Madsen

REFERENCES

- Carroll, J. M. (1997). Human-computer interaction: Psychology as a science of design. *Annual Review of Psychology*, 48, 61–83.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge: MIT Press.
- Newell, A., & Simon, H. A. (1961). Computer simulation of human thinking. *Science*, 134, 2011–2017.
- Newell, A., & Simon, H. A. (1971). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

PROXY REPORTING

Proxy reporting refers to two related situations in conducting social surveys. If, by certain preset criteria, survey researchers consider a sample person as being unable to answer the survey, then the proxy or substitute respondent is used. More generally, the situation

in which a person who reports information on other persons can be defined as proxy reporting.

—Tim Futing Liao

PROXY VARIABLE

A proxy variable is a variable that is used to measure an unobservable quantity of interest. Although a proxy variable is not a direct measure of the desired quantity, a good proxy variable is strongly related to the unobserved variable of interest. Proxy variables are extremely important to and frequently used in the social sciences because of the difficulty or impossibility of obtaining measures of the quantities of interest.

An example clarifies both the importance and the problems resulting from the use of proxy variables. Suppose a researcher is interested in the effect of intelligence on income, controlling for age. If it were possible to know the intelligence of an individual and assuming (for purposes of illustration) that the relationship is linear, then the relationship could be estimated using MULTIPLE REGRESSION by

$$\text{Income}_i = \alpha + \beta_1 \cdot \text{Intelligence}_i + \beta_2 \cdot \text{age}_i + \varepsilon_i.$$

Unlike age, intelligence is not a quantity that can be directly measured. Instead, the researcher interested in the above relationship must rely on measures arguably related to intelligence such as years of education and standardized test results. Using years of education as a proxy variable for intelligence assumes: $\text{Intelligence}_i = \text{Years of Education}_i + \delta_i$. Consequently, the researcher estimates the following:

$$\text{Income}_i = \alpha + \beta_1 \cdot \text{Years of Education}_i + \beta_2 \cdot \text{age}_i + \varepsilon_i.$$

Strictly speaking, β_1 does not measure the covariation between intelligence and income unless $\delta_i = 0$ for all individuals i . However, because it is impossible to assess how well a proxy variable measures the unobservable quantity of interest (i.e., what the value of δ_i is), and because the relationship between income and years of education is of interest only insofar as it permits an understanding of the relationship between income and intelligence, β_1 is frequently interpreted in terms of the latter. Some suggest that when several measures are available (as is arguably the case in the above example), specification searches may be used to try to select the best proxy variable (Leamer, 1978).

Statistically, the decision of whether to use a proxy variable entails weighing two possible consequences. If a proxy variable is used, then the researcher necessarily encounters the traditional problems associated with “errors-in-variables” (i.e., all PARAMETER ESTIMATES are inconsistent) (Fuller, 1987). However, if no proxy variable is used and the unmeasured quantity is related to the specification of interest, then failure to use a proxy variable will result in omitted variable bias (see OMITTED VARIABLE and BIAS). Econometric investigation of the trade-off reveals that it is generally advisable to include a proxy variable, even one that is relatively poor (Kinal & Lahiri, 1983).

—Joshua D. Clinton

REFERENCES

- Fuller, W. A. (1987). *Measurement error models*. New York: John Wiley.
- Kinal, T., & Lahiri, K. (1983). Specification error analysis with stochastic regressors. *Econometrica*, 54, 1209–1220.
- Leamer, E. (1978). *Specification searches: Ad hoc inferences with nonexperimental data*. New York: John Wiley.
- Maddala, G. S. (1992). *Introduction to econometrics*. Englewood Cliffs, NJ: Prentice Hall.

PSEUDO-R-SQUARED

A commonly used GOODNESS-OF-FIT MEASURE in REGRESSION analysis is the COEFFICIENT OF (MULTIPLE) DETERMINATION, also known as *R-SQUARED* or R^2 . R^2 essentially measures the proportion of the variance in the DEPENDENT VARIABLE that is explained by the MULTIPLE REGRESSION model. Although some disagreement exists as to the overall utility of R^2 , it does provide a widely used measure of model performance.

Regression analysts have available a wide array of diagnostic tools and goodness-of-fit measures. Such tools and measures are less well developed for models using PROBIT and LOGIT. In particular, there is no direct analog to R^2 as a goodness-of-fit measure. Several substitutes, pseudo- R^2 s, are available and are often present in the results of various computer programs. A difficulty facing probit and logit analysts is in choosing among the pseudo- R^2 s. An initial question to be addressed when choosing a pseudo- R^2 will be the use to which the measure will be put (e.g., explanation of VARIANCE, HYPOTHESIS testing, classifying the dependent variable).

The most common use is as a measure of how much variation in the dependent variable is explained or accounted for by the model. One frequently used method, developed by McKelvey and Zavoina (1975), is to use the estimated probit or logit coefficients to compute the variance of the forecasted values for the LATENT DEPENDENT VARIABLE, denoted $\text{var}(\hat{y}_i)$. The disturbances have unit variance, and the total variance then reduces to explained variance plus 1. This pseudo- R^2 is then calculated as

$$R^2 = \frac{\text{var}(\hat{y}_i)}{1 + \text{var}(\hat{y}_i)}.$$

Another popular alternative pseudo- R^2 , developed by Aldrich and Nelson (1984), uses the chi-square statistic, $-2LLR$, defined as $-2 \ln(L_0/L_1)$ with k , the number of independent variables estimated, DEGREES OF FREEDOM. This measure requires two passes through the data—the first with only the constant to determine L_0 and the second with the full model to determine L_1 . The pseudo- R^2 is then calculated as

$$R^2 = \frac{-2LLR}{N - 2LLR},$$

where N is the number of observations.

Both measures proved to be good estimates of the ORDINARY LEAST SQUARES R^2 in a test conducted by Hagle and Mitchell (1992) using SIMULATION data. The Aldrich and Nelson (1984) measure produced slightly smaller errors and was more ROBUST when the assumption of a NORMAL DISTRIBUTION in the dependent variable was violated. A theoretical limitation in the calculation of the Aldrich and Nelson measure is easily corrected based on the percentage of observations in the modal category.

The utility of pseudo- R^2 s for probit and logit is based on the assumption of an unobserved continuous variable that manifests itself only as an observed dichotomous variable. For example, the unobserved variable may be the probability that a voter will choose Candidate A over Candidate B, and the observed variable is the actual choice between A and B. Because variations in the unobserved variable will produce the same observed variable, analysts may place greater emphasis on the model's ability to correctly classify the observations. One issue in using the correct classification rate as a goodness-of-fit measure is that it often becomes difficult for probit and logit to correctly classify observations into the nonmodal category when

the percentage of observations in the modal category is quite high. On the whole, a combination of different measures may provide the best understanding of model performance.

—Timothy M. Hagle

REFERENCES

- Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit and probit models*. Thousand Oaks, CA: Sage.
- Hagle, T. M., & Mitchell, G. E., II. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 36, 762–784.
- McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, 103–120.

PSYCHOANALYTIC METHODS

Psychoanalysis is both a clinical psychotherapeutic practice and a body of theory concerning people's mental life, including its development and expression in activities and relationships. Empirical findings are basic to psychoanalysis, but rather than being based on a claim to objectivity, these findings are recognized to be the product of both the patient's and analyst's subjectivities. Objections to the scientific status of psychoanalytic data have been on the grounds that the psychoanalytic method was based in the consulting room, not the laboratory, and produced single case studies, not generalizable findings. However, following the challenge to positivist methods in social science research, the resulting linguistic, qualitative, and hermeneutic turns have opened a space for psychoanalysis in qualitative methods.

Psychoanalytic theory has proved influential in questioning the Enlightenment subject—the rational, unitary, consciously intentional subject that usually underpins social scientists' assumptions about research participants. The premise of unconscious conflict and defenses casts doubt on the idea of a self-knowable, transparent respondent, an idea underpinning interview-based, survey, and attitude scale research (Hollway & Jefferson, 2000). Such unconscious contents necessarily require interpretation, and these are warranted by psychoanalytic concepts (e.g., the idea of denying an action when its recall would be painful). Such concepts have been built up and refined through clinical psychoanalytic practice (Dreher, 2000).

Psychoanalytic methods require modification if extended beyond the consulting room, where research findings are subordinate to therapeutic purpose. Clinical and research applications both use psychoanalytic interpretation. In the consulting room, analysts' interpretations are a small part of their therapeutic method, offered directly to patients whose emotional responses provide a test of the interpretation's correctness. In generating and interpreting research data psychoanalytically, other methods of validation are sought (e.g., looking for supporting and countervailing examples in the whole text or detecting a recognizable pattern of defenses).

The method of free association, on which psychoanalysis is based, encourages patients to report their thoughts without reservation on the assumption that uncensored thoughts lead to what is significant. The use of free association can inform the production and analysis of interview data (see FREE ASSOCIATION NARRATIVE INTERVIEW METHOD), tapping aspects of human social life of which methods relying on the self-knowledgeable transparent research subject remain unaware (e.g., the desires and anxieties that affect family and work relationships). The concept of transference refers to the displacement of unconscious wishes from an earlier figure, such as a parent, on to the analyst, where they are experienced with the immediacy of feelings for the current person. Albeit diluted in a nonclinical setting, researchers sensitive to their own and participants' transferences and countertransferences can use these as data (e.g., Walkerdine, Lucey, & Melody, 2001). Epistemologically, these concepts enhance the critique of objectivity and supplement the understanding of researcher reflexivity, an important but undertheorized concept in qualitative social science.

Interviewing and observation both include psychoanalytic influences. Kvale (1999) argues that knowledge produced in the psychoanalytic interview is relevant for research. Two currents of observation work use psychoanalytic methods—namely, ethnographic field studies (Hunt, 1989) and baby observation (Miller, Rustin, Rustin, & Shuttleworth, 1989). In Germany, ethnopsychanalysis applies psychoanalysis to the study of culture. Psychoanalytic observation—dating from the 1940s in England, when detailed once-weekly observation of babies in their families became a part of training at the Tavistock Institute—now extends to research on wider topics.

—Wendy Hollway

REFERENCES

- Dreher, A. U. (2000). *Foundations for conceptual research in psychoanalysis* (Psychoanalytic Monograph 4). London: Karnac.
- Hollway, W., & Jefferson, T. (2000). *Doing qualitative research differently: Free association, narrative and the interview method*. London: Sage.
- Hunt, J. C. (1989). *Psychoanalytic aspects of fieldwork* (University Paper Series on Qualitative Research Methods 18). Newbury Park, CA: Sage.
- Kvale, S. (1999). The psychoanalytic interview as qualitative research. *Qualitative Inquiry*, 5(1), 87–113.
- Miller, L., Rustin, M., Rustin, M., & Shuttleworth, J. (1989). *Closely observed infants*. London: Duckworth.
- Walkerdine, V., Lucey, H., & Melody, J. (2001). *Growing up girl: Psychosocial explorations of gender and class*. Basingstoke, UK: Palgrave.

PSYCHOMETRICS

Psychometrics deals with the measurement of individual differences between people. The instruments used for measurement can be psychological tests, questionnaires, personality inventories, and observational checklists. The measurement values quantify either the levels of a psychological property or the ordering of these levels. These properties can be abilities (e.g., verbal intelligence, analytical reasoning, spatial orientation), skills (e.g., in manipulating control panels for traffic regulation, in chairing a meeting in a work situation), personality traits (e.g., introversion, friendliness, neuroticism), preferences (e.g., for consumer goods, politicians, authors), and attitudes (e.g., toward abortion, the political right wing, NATO intervention in other countries). “John is more introvert than Mary” is an example of an ordering with respect to a personality trait, and “Carol’s general intelligence is high enough to admit her to the course” compares Carol’s general cognitive ability level to the absolute level required to pass the course entrance.

The measurement values obtained through tests, questionnaires, or other instruments must be reliable and valid. A measurement is reliable to the degree it can be generalized across independent replications of the same measurement procedure. A broader definition says that generalization can also be across the measurement units (the items from a test), different measurement occasions (time points), or different psychologists who assessed a pupil or a client.

A measurement has CONSTRUCT VALIDITY to the degree it is based on an adequate operationalization of the underlying property. This means that the measurement instrument measures introversion and not, say, the respondent's idealized self-image. Predictive validity holds to the degree the measurement permits making correct predictions about individuals' future behavior (e.g., the degree to which analytical reasoning predicts people's suitability for computer programmer).

Psychometrics studies and develops mathematical and statistical measurement methods. Older psychometric methods are classical test theory, mostly due to Charles Spearman; LIKERT SCALING for rating scale data; FACTOR ANALYSIS as developed in the 1920s and 1930s by Spearman, Thurstone, and Thompson; Thurstone's method for paired comparison data; Guttman's scalogram analysis for dominance data; and Coombs's unfolding method for preference data. GENERALIZABILITY THEORY, mostly due to Cronbach, broadens classical test theory by using an analysis of variance layout of sources that explain test performance. ITEM RESPONSE THEORY models are a large family of modern psychometric models, initially developed in the 1950s and 1960s by Lord, Birnbaum, Rasch, and Mokken. Since then, this development has been taken up by many researchers and has dominated psychometrics over the past few decades.

Traditionally, MULTIVARIATE ANALYSIS methods have been part of psychometrics. In particular, factor analysis and CLUSTER ANALYSIS have been used for the investigation of construct validity and regression methods for establishing test batteries to be used for prediction. Methods for psychophysical measurement, such as the determination of visual or aural thresholds, are mainly aimed at establishing general laws of behavior and less at measuring individual differences.

Psychometrics also includes the technology of constructing measurement instruments of all sorts, including standard paper-and-pencil tests and item banks for computerized adaptive testing procedures that enable a more efficient measurement procedure. Other applications identify items that put certain minority groups at a disadvantage (differential item functioning) and respondents who produced aberrant patterns of item scores (person-fit analysis). Also included are applications that study the cognitive processes and solution strategies respondents use to solve the items in the test (cognitive modeling).

—Klaas Sijtsma

See also ITEM RESPONSE THEORY

PSYCHOPHYSIOLOGICAL MEASURES

Psychophysiological measures are physiological measures used to index psychological constructs (e.g., psychological states or processes). Researchers began to explore mind-body relationships scientifically early in the 20th century. However, psychophysiology did not emerge as a subdiscipline separate from physiological psychology until mid-century. Until that time (and beyond), physiological psychologists explored the effects of physiological changes, particularly neurophysiological ones, on behavioral and mental outcome measures, typically conducting animal experiments. In contrast, psychophysiolgists explored effects of psychological changes on physiological outcome measures, typically conducting human experiments. Now, the distinctions between psychophysiology and physiological psychology have blurred, with both subsumed within neuroscience more generally.

Nevertheless, psychophysiolgists have made enduring general contributions to psychology and behavioral science as well as contributing many important and useful physiological measures of indexes of psychological states. The utility of physiological indexes of psychological states rests on the assumption that the observation and recording of biological or physiological processes, particularly less consciously controllable ones, are methodologically advantageous.

Physiological and, hence, psychophysiological measures are typically online, covert, and continuous. Online physiological measures provide pinpoint temporal accuracy. They can be linked in time to the expected occurrence of the psychological state or process being indexed, thereby providing simultaneous evidence of the strength or operation of the state or process, while avoiding some pitfalls of post hoc, measures such variations in memory. For example, an online physiological measure of happiness would avoid any measurement error due to participant recall.

Covert physiological measures provide data not readily observable either to the participant being assessed or to outside observers, thereby increasing their veridicality. Given their covert nature, physiological measures provide no conscious feedback to

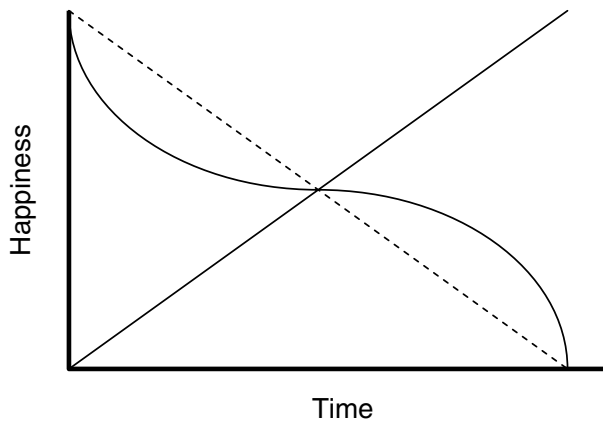


Figure 1 Psychophysiological Measures

participants, thereby providing little if any basis for participant control of their responses. Hence, by their covert nature, physiological measures void participant impression management strategies in experimental contexts. For example, a covert measure of happiness would void attempts of participants to appear more or less happy than they actually were.

Continuous physiological measures provide topographical accuracy. Experimenters can sample physiological measures repeatedly at high rates. Indeed, given computer technology, physiological measures can be recorded with near-perfect fidelity over time. Hence, researchers can produce topographical physiological records that increase available information without increasing cost. For example, a topographical record of happiness would allow investigators to examine trends and variability in happiness within an epoch. Even if measurements share CENTRAL TENDENCY values (i.e., means, medians, modes), topographical differences can provide important psychological information. Figure 1 illustrates a few of the many possible topographical differences among participant responses sharing the same central tendencies.

However, the fact that physiological measures typically provide online, covert, and continuous data, with corresponding temporal, veridical, and topographical accuracy, does not mean they have CONSTRUCT VALIDITY. Here, the legacy of psychophysiology for methodologists is less obvious. Indeed, the happiness examples above mention no specific physiological index or metric. Even though a literature search of happiness studies would probably turn up quite a few

physiological measures that have been used, such as electrodermal activity, heart rate, blood flow to the cerebral cortex, and so forth, explicit theory linking any of these or other physiological measures to the happiness construct might be quite fuzzy.

Physiological measures, like other measures used in psychological research, must have construct validity. As stated elsewhere (Blascovich, 2000; Cacioppo, Tassinary, & Berntson, 2000), *invariance* is the ideal association between a psychological construct and its index, whether physiological or not. This means that the construct and its index should co-occur, with the presence of one ensuring the presence of the other and vice versa. However, invariance proves a difficult ideal to achieve between psychological constructs and physiological indexes because the constructs (e.g., happiness, love, risk taking, prejudice, learning) are difficult to define explicitly, the body operates more as a general-purpose system than a specific purpose one, or both.

Here, as for nonphysiological measures, one needs a specific measurement theory on which to base any invariance assumption. For physiological measures, the measurement theory must be based on psychophysiological theory connecting the construct with its purported index. This is no mean feat. Nevertheless, several strong theories have been applied or developed to provide the necessary justification for many psychological constructs, including ones within many psychological domains such as affect, motivation, and cognition (for reviews, see Blascovich, 2000; Cacioppo, Tassinary, & Berntson, 2000).

For example, Cacioppo, Petty, Losch, and Kim (1986) suggested using patterns of zygomaticus major and corrugator supercilii (facial) muscle responses assessed via electromyographic (EMG) techniques to index and distinguish positive and negative affect. The psychophysiological basis for their assertion involved Darwinian work on the evolutionary significance of facial expressions for social interactions. They reasoned that positive affect should be accompanied by increases in zygomaticus major (the “smile muscle”) EMG and decreases in corrugator supercilii (the “frown muscle”) EMG but that negative affect should be accompanied by the reverse pattern. Subsequently, they validated these patterns experimentally.

Blascovich and his colleagues (e.g., Blascovich & Tomaka, 1996) suggested using patterns of cardiovascular responses assessed using electrocardiographic (ECG), impedance cardiographic (ZKG), and

hemodynamic blood pressure monitoring techniques to index challenge and threat motivational states. The psychophysiological basis for their assertion involved Dienstbier's (1989) neuroendocrine theory of physiological toughness. Blascovich and colleagues demonstrated that challenge, which they identified as a benign motivational state (i.e., when resources are sufficient to meet situational demands), is accompanied by sympathetic adrenal medullary (SAM)-driven increases in cardiac contractility and cardiac output accompanied by vasodilation, whereas threat, which they identified as a malignant motivational state (i.e., when resources are insufficient to meet situational demands), is accompanied by pituitary adrenal cortical (PAC) inhibition of the SAM axis, resulting in increases in cardiac contractility but little change in cardiac output and vasomotor tone.

Many other theoretically based physiological indexes of psychological constructs have appeared and continue to add to the store of such measures. With new technology, such as brain imaging techniques, constantly emerging and increased appreciation of the utility of physiological measures for psychological research, we can look forward to even more and better ones being developed.

—Jim Blascovich

REFERENCES

- Blascovich, J. (2000). Psychophysiological indexes of psychological processes. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 117–137). Cambridge, UK: Cambridge University Press.
- Blascovich, J., & Tomaka, J. (1996). The biopsychosocial model of arousal regulation. In M. Zanna (Ed.), *Advances in experimental social psychology* (pp. 1–51). New York: Academic Press.
- Cacioppo, J. T., Petty, R. E., Losch, M. E., & Kim, H. S. (1986). Electromyographic specificity during simple physical and attitudinal tasks: Location and topographical features of integrated EMG responses. *Biological Psychology*, 18, 85–121.
- Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2000). *Handbook of psychophysiology* (2nd ed.). Cambridge, UK: Cambridge University Press.
- Dienstbier, R. A. (1989). Arousal and physiological toughness: Implications for mental and physical health. *Psychological Review*, 96, 84–100.

PUBLIC OPINION RESEARCH

Public opinion research involves the CONCEPTUALIZATION and MEASUREMENT of what people think or how they behave. Since the development of the field in conjunction with rise of survey research, practitioners have disagreed about the meaning and measurement of each of these aspects, and research on “public opinion research” remains an active and contested area today. Although an argument can be made that “public opinion” can be represented as well by various forms of collective action, such as strikes and protest marches, as it can by survey responses, SURVEYS and polls remain the predominant basis for measuring and reporting it.

One of the key conceptual points is whether public opinion is an individual or an AGGREGATED concept. Although an individual may have or express attitudes and opinions, does a public exist only as the aggregated combination of the views of individuals? If individuals hold views but they are not measured, aggregated, and communicated to others, is there in fact any prospect of collective action based on such opinions? Another important point is whether public opinion exists if it is not expressed. This argument has been used to distinguish between “measured” opinion and “considered” opinion. This raises the prospect that anything that can be measured with a survey can be considered “opinion” no matter what the quality of the questions asked, the definition of the groups that were queried, or how much thought an individual has given to an issue. *Measured opinion* is reflected in the responses that individuals give to questions they are asked, whether or not they have devoted any attention to the topic, whereas *considered opinion* is the end result of a process of deliberation that takes place over time.

Generally, public opinion is associated with the results from a survey, typically involving measurements taken from a REPRESENTATIVE SAMPLE of a relevant POPULATION. Typically, this would be designed to reflect the attitudes of all adults in a society or political constituency. Some researchers would argue that there are other behavioral measures of public opinion, such as voting, demonstrations, or letters that are written to a newspaper.

The study of public opinion has focused on two key areas: the conceptual dimensions of the opinions themselves and the effects of survey methodology on

the views expressed. Most populations are stratified in opinion terms, with a very small proportion comprising political elites who hold complex, well-structured views on issues and who can provide survey responses that seem to be ideologically consistent and constrained (Converse, 1964). Research shows that elites, in addition to their discourse through the media, provide significant cues to the mass public about the salience or importance of specific issues, as well as the specific domains along which they should be considered. Information about the party affiliation or political ideology of elites can be very useful to citizens when evaluating such reports, helping them to make judgments without a significant investment in learning or assimilating new information (Lupia, McCubbins, & Popkin, 1998; Popkin, 1991). But information is not the only basis for holding opinions; other social psychological concepts such as stereotypes may support their formation. This is especially true for attitudes about race and gender, for example, or about individuals on welfare or in poverty.

Another aspect of opinions is the intensity with which the attitudes are held. In the political arena, strength of feelings among a minority may be more powerful in explaining policy or legislative changes than the distribution of opinions in the entire population. This can be especially important if such intensity is related to forms of political behavior such as protest, letter writing, or voting.

From a methodological perspective, public opinion researchers are concerned that survey respondents often answer questions posed to them as a matter of courtesy or an implied social contract with an interviewer. As a result, they express opinions that may be lightly held or come off the top of their heads in response to a specific question or its wording. The use of CLOSED-ENDED QUESTIONS, especially those that do not provide an explicit alternative such as "Don't know" or "Have not thought about that," can produce measured opinions that are weakly held at best.

Increasingly, other methodological issues may relate to potential BIASES in the selection of those whose opinions are measured. They include problems of self-selected respondents who answer questions on the growing number of Web sites that simulate polls, problems of the increased use of cell phones that are generally unavailable in the standard RANDOM DIGIT DIALING (RDD) sample, and declining response rates.

Because of concerns about the meaning of public opinion and the ways that survey methodology can

affect how it is measured, public opinion research will remain an active area of substantive and methodological attention.

—Michael W. Traugott

REFERENCES

- Converse, P. E. (1964). The nature of belief systems in mass publics. In D. Apter (Ed.), *Ideology and discontent* (pp. 206–261). New York: Free Press.
- Lupia, A., McCubbins, M. D., & Popkin, S. L. (Eds.). (1998). *Elements of reason*. Cambridge, UK: Cambridge University Press.
- Popkin, S. L. (1991). *The reasoning voter*. Chicago: University of Chicago Press.
- Price, V. (1992). *Public opinion*. Newbury Park, CA: Sage.
- Zaller, J. R. (1992). *The nature and origins of mass opinion*. Cambridge, UK: Cambridge University Press.

PURPOSIVE SAMPLING

Purposive sampling in qualitative inquiry is the deliberate seeking out of participants with particular characteristics, according to the needs of the developing analysis and emerging theory. Because, at the beginning of the study, the researcher does not know enough about a particular phenomenon, the nature of the sample is not always predetermined. Often, midway through data analysis, the researcher will realize that he or she will need to interview participants who were not envisioned as pertinent for the study at the beginning of the project. For example, Martin (1998), in her study of sudden infant death syndrome (SIDS), discovered issues of control over the responsibility for these infants between the groups of first responders (emergency medical technicians and the firefighters) and realized that she would have to interview these professionals, extending her study beyond the interviews with parents, as originally perceived.

Types of purposive sampling are nominated or SNOWBALL SAMPLING (in which participants are referred by members of the same group who have already been enrolled in the study) and THEORETICAL SAMPLING (in which participants are deliberately sought according to information required by the analysis as the study progresses).

In nominated or snowball sampling, the researcher locates a "good" participant and, at the end of the interview, asks the participant to help with the study

by referring the researcher to another person who may like to participate in the study. Thus, sampling follows a social network. Nominated or snowball sampling is particularly useful when groups are hard to identify or may not volunteer or respond to a notice advertising for participants. It is a useful strategy when locating participants who would otherwise be hard to locate, perhaps because of shame or fear of reprisal for illegal activities, such as drug use; a closed group, such as a motorcycle gang; or those who have private behaviors or a stigma associated with a disease. Gaining trust with the first participant and allowing that person to assure the group that the research is "OK" provides access to participants who would otherwise be unobtainable.

The final use of a nominated sample is by researchers who are using theoretical sampling. They may use a form of nominated sample by requesting recommendations from the group for participants who have certain kind of experiences or knowledge needed to move the analysis forward.

—Janice M. Morse

REFERENCE

- Martin, K. (1998). *When a baby dies of SIDS: The parents' grief and searching for a reason*. Edmonton, Canada: Qual Institute Press.

PYGMALION EFFECT

The term *Pygmalion effect* has both a general and a specific meaning. Its general meaning is that of an *interpersonal expectancy effect* or an *interpersonal self-fulfilling prophecy*. These terms refer to the research finding that the behavior that one person expects from another person is often the behavior that is, in fact, obtained from that other person. The more specific meaning of the *Pygmalion effect* is the effect of teachers' expectations on the performance of their students.

The earliest EXPERIMENTAL studies of Pygmalion effects were conducted with human research participants. Experimenters obtained ratings of photographs of stimulus persons from their research participants, but half the experimenters were led to expect high photo ratings and half were led to expect low photo ratings. In a series of such studies, experimenters

expecting higher photo ratings obtained substantially higher photo ratings than did experimenters expecting lower photo ratings (Rosenthal, 1976).

To investigate the generality of these interpersonal expectancy effects in the laboratory, researchers conducted two studies employing animal subjects (Rosenthal, 1976). Half the experimenters were told their rats had been specially bred for maze (or Skinner box) brightness, and half were told their rats had been specially bred for maze (or Skinner box) dullness. In both experiments, when experimenters had been led to expect better learning from their rat subjects, they obtained better learning from their rat subjects.

The "Pygmalion experiment," as it came to be known, was a fairly direct outgrowth of the studies just described. If rats became brighter when expected to by their experimenters, then it did not seem far-fetched to think that children could become brighter when expected to by their teacher. Accordingly, all of the children in the study (Rosenthal & Jacobson, 1968, 1992) were administered a nonverbal test of intelligence, which was disguised as a test that would predict intellectual "blooming." There were 18 classrooms in the school, 3 at each of the six grade levels. Within each grade level, the 3 classrooms were composed of children with above-average ability, average ability, and below-average ability, respectively. Within each of the 18 classrooms, approximately 20% of the children were chosen at random to form the experimental group. Each teacher was given the names of the children from his or her class who were in the experimental condition. The teacher was told that these children had scored on the predictive test of intellectual blooming such that they would show surprising gains in intellectual competence during the next 8 months of school. The only difference between the experimental group and the control group children, then, was in the mind of the teacher.

At the end of the school year, 8 months later, all the children were retested with the same test of intelligence. Considering the school as a whole, the children from whom the teachers had been led to expect greater intellectual gain showed a significantly greater gain than did the children of the control group.

There are now nearly 500 experiments investigating the phenomenon of interpersonal expectancy effects or the Pygmalion effect broadly defined. The typical obtained magnitude of the effect is very substantial; when we have been led to expect a certain behavior from others, we are likely to obtain it about 65% of

the time, compared to about 35% of the time, when we have not been led to expect that behavior.

—Robert Rosenthal

See also EXPERIMENTER EXPECTANCY EFFECT,
INVESTIGATOR EFFECTS

REFERENCES

- Blanck, P. D. (Ed.). (1993). *Interpersonal expectations: Theory, research, and applications*. New York: Cambridge University Press.
- Eden, D. (1990). *Pygmalion in management: Productivity as a self-fulfilling prophecy*. Lexington, MA: D. C. Heath.
- Rosenthal, R. (1976). *Experimenter effects in behavioral research: Enlarged edition*. New York: Irvington.
- Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston.
- Rosenthal, R., & Jacobson, L. (1992). *Pygmalion in the classroom: Expanded edition*. New York: Irvington.

The SAGE Encyclopedia of
**SOCIAL SCIENCE
RESEARCH METHODS**

GENERAL EDITORS

Michael S. Lewis-Beck

Alan Bryman

Tim Futing Liao

EDITORIAL BOARD

Leona S. Aiken

Arizona State University, Tempe

R. Michael Alvarez

California Institute of Technology, Pasadena

Nathaniel Beck

University of California, San Diego, and New York University

Norman Blaikie

*Formerly at the University of Science,
Malaysia and the RMIT University, Australia*

Linda B. Bourque

University of California, Los Angeles

Marilynn B. Brewer

Ohio State University, Columbus

Derek Edwards

Loughborough University, United Kingdom

Floyd J. Fowler, Jr.

University of Massachusetts, Boston

Martyn Hammersley

The Open University, United Kingdom

Guillermina Jasso

New York University

David W. Kaplan

University of Delaware, Newark

Mary Maynard

University of York, United Kingdom

Janice M. Morse

University of Alberta, Canada

Martin Paldam

University of Aarhus, Denmark

Michael Quinn Patton

*The Union Institute & University,
Utilization-Focused Evaluation, Minnesota*

Ken Plummer

University of Essex, United Kingdom

Charles Tilly

Columbia University, New York

Jeroen K. Vermunt

Tilburg University, The Netherlands

Kazuo Yamaguchi

University of Chicago, Illinois

The SAGE Encyclopedia of
**SOCIAL SCIENCE
RESEARCH METHODS**

VOLUME 3

MICHAEL S. LEWIS-BECK

Department of Political Science, University of Iowa

ALAN BRYMAN

Department of Social Sciences, Loughborough University

TIM FUTING LIAO

Department of Sociology, University of Essex and University of Illinois, Urbana-Champaign

EDITORS

A SAGE Reference Publication

 **SAGE Publications**
International Educational and Professional Publisher
Thousand Oaks ■ London ■ New Delhi

Copyright © 2004 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

For information:



SAGE Publications, Inc.
2455, Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
6, Bonhill Street
London EC2A 4PU
United Kingdom

SAGE Publications India Pvt. Ltd.
B-42, Panchsheel Enclave
Post Box 4109
New Delhi 110 017 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Lewis-Beck, Michael S.

The SAGE encyclopedia of social science research methods/Michael S.

Lewis-Beck, Alan Bryman, Tim Futing Liao.

p. cm.

Includes bibliographical references and index.

ISBN 0-7619-2363-2 (cloth)

1. Social sciences—Research—Encyclopedias. 2. Social sciences—Methodology—Encyclopedias.

I. Bryman, Alan. II. Liao, Tim Futing. III. Title.

H62.L456 2004

300!.! 72—dc22

2003015882

03 04 05 06 10 9 8 7 6 5 4 3 2 1

<i>Acquisitions Editor:</i>	Chris Rojek
<i>Publisher:</i>	Rolf A. Janke
<i>Developmental Editor:</i>	Eileen Haddox
<i>Editorial Assistant:</i>	Sara Tauber
<i>Production Editor:</i>	Melanie Birdsall
<i>Copy Editors:</i>	Gillian Dickens, Liann Lech, A. J. Sobczak
<i>Typesetter:</i>	C&M Digital (P) Ltd.
<i>Proofreader:</i>	Mattson Publishing Services, LLC
<i>Indexer:</i>	Sheila Bodell
<i>Cover Designer:</i>	Ravi Balasuriya

List of Entries

Abduction
Access
Accounts
Action Research
Active Interview
Adjusted *R*-Squared. *See* Goodness-of-Fit Measures; Multiple Correlation; *R*-Squared
Adjustment
Aggregation
AIC. *See* Goodness-of-Fit Measures
Algorithm
Alpha, Significance Level of a Test
Alternative Hypothesis
AMOS (Analysis of Moment Structures)
Analysis of Covariance (ANCOVA)
Analysis of Variance (ANOVA)
Analytic Induction
Anonymity
Applied Qualitative Research
Applied Research
Archival Research
Archiving Qualitative Data
ARIMA
Arrow's Impossibility Theorem
Artifacts in Research Process
Artificial Intelligence
Artificial Neural Network. *See* Neural Network
Association
Association Model
Assumptions
Asymmetric Measures
Asymptotic Properties
ATLAS.ti
Attenuation
Attitude Measurement
Attribute
Attrition
Auditing
Authenticity Criteria
Autobiography
Autocorrelation. *See* Serial Correlation
Autoethnography
Autoregression. *See* Serial Correlation
Average
Bar Graph
Baseline
Basic Research
Bayes Factor
Bayes' Theorem, Bayes' Rule
Bayesian Inference
Bayesian Simulation. *See* Markov Chain Monte Carlo Methods
Behavior Coding
Behavioral Sciences
Behaviorism
Bell-Shaped Curve
Bernoulli
Best Linear Unbiased Estimator
Beta
Between-Group Differences
Between-Group Sum of Squares
Bias
BIC. *See* Goodness-of-Fit Measures
Bimodal
Binary
Binomial Distribution
Binomial Test
Biographic Narrative Interpretive Method (BNIM)
Biographical Method. *See* Life History Method
Biplots
Bipolar Scale
Biserial Correlation
Bivariate Analysis
Bivariate Regression. *See* Simple Correlation (Regression)
Block Design
BLUE. *See* Best Linear Unbiased Estimator
BMDP
Bonferroni Technique

- Bootstrapping
Bounded Rationality
Box-and-Whisker Plot. *See* Boxplot
Box-Jenkins Modeling
Boxplot

CAIC. *See* Goodness-of-Fit Measures
Canonical Correlation Analysis
CAPI. *See* Computer-Assisted Data Collection
Capture-Recapture
CAQDAS (Computer-Assisted Qualitative Data Analysis Software)
Carryover Effects
CART (Classification and Regression Trees)
Cartesian Coordinates
Case
Case Control Study
Case Study
CASI. *See* Computer-Assisted Data Collection
Catastrophe Theory
Categorical
Categorical Data Analysis
CATI. *See* Computer-Assisted Data Collection
Causal Mechanisms
Causal Modeling
Causality
CCA. *See* Canonical Correlation Analysis
Ceiling Effect
Cell
Censored Data
Censoring and Truncation
Census
Census Adjustment
Central Limit Theorem
Central Tendency. *See* Measures of Central Tendency
Centroid Method
Ceteris Paribus
CHAID
Change Scores
Chaos Theory
Chi-Square Distribution
Chi-Square Test
Chow Test
Classical Inference
Classification Trees. *See* CART
Clinical Research
Closed Question. *See* Closed-Ended Questions
Closed-Ended Questions
Cluster Analysis
Cluster Sampling
Cochran's Q Test
Code. *See* Coding
Codebook. *See* Coding Frame
Coding
Coding Frame
Coding Qualitative Data
Coefficient
Coefficient of Determination
Cognitive Anthropology
Cohort Analysis
Cointegration
Collinearity. *See* Multicollinearity
Column Association. *See* Association Model
Commonality Analysis
Communality
Comparative Method
Comparative Research
Comparison Group
Competing Risks
Complex Data Sets
Componential Analysis
Computer Simulation
Computer-Assisted Data Collection
Computer-Assisted Personal Interviewing
Concept. *See* Conceptualization, Operationalization, and Measurement
Conceptualization, Operationalization, and Measurement
Conditional Likelihood Ratio Test
Conditional Logit Model
Conditional Maximum Likelihood Estimation
Confidence Interval
Confidentiality
Confirmatory Factor Analysis
Confounding
Consequential Validity
Consistency. *See* Parameter Estimation
Constant
Constant Comparison
Construct
Construct Validity
Constructionism, Social
Constructivism
Content Analysis
Context Effects. *See* Order Effects
Contextual Effects
Contingency Coefficient. *See* Symmetric Measures
Contingency Table
Contingent Effects

-
- Continuous Variable
 Contrast Coding
 Control
 Control Group
 Convenience Sample
 Conversation Analysis
 Correlation
 Correspondence Analysis
 Counterbalancing
 Counterfactual
 Covariance
 Covariance Structure
 Covariate
 Cover Story
 Covert Research
 Cramér's *V*. *See* Symmetric Measures
 Creative Analytical Practice (CAP) Ethnography
 Cressie-Read Statistic
 Criterion Related Validity
 Critical Discourse Analysis
 Critical Ethnography
 Critical Hermeneutics
 Critical Incident Technique
 Critical Pragmatism
 Critical Race Theory
 Critical Realism
 Critical Theory
 Critical Values
 Cronbach's Alpha
 Cross-Cultural Research
 Cross-Lagged
 Cross-Sectional Data
 Cross-Sectional Design
 Cross-Tabulation
 Cumulative Frequency Polygon
 Curvilinearity

 Danger in Research
 Data
 Data Archives
 Data Management
 Debriefing
 Deception
 Decile Range
 Deconstructionism
 Deduction
 Degrees of Freedom
 Deletion
 Delphi Technique
 Demand Characteristics

 Democratic Research and Evaluation
 Demographic Methods
 Dependent Interviewing
 Dependent Observations
 Dependent Variable
 Dependent Variable (in Experimental Research)
 Dependent Variable (in Nonexperimental Research)
 Descriptive Statistics. *See* Univariate Analysis
 Design Effects
 Determinant
 Determination, Coefficient of. *See* *R*-Squared
 Determinism
 Deterministic Model
 Detrending
 Deviant Case Analysis
 Deviation
 Diachronic
 Diary
 Dichotomous Variables
 Difference of Means
 Difference of Proportions
 Difference Scores
 Dimension
 Disaggregation
 Discourse Analysis
 Discrete
 Discriminant Analysis
 Discriminant Validity
 Disordinality
 Dispersion
 Dissimilarity
 Distribution
 Distribution-Free Statistics
 Disturbance Term
 Documents of Life
 Documents, Types of
 Double-Blind Procedure
 Dual Scaling
 Dummy Variable
 Duration Analysis. *See* Event History Analysis
 Durbin-Watson Statistic
 Dyadic Analysis
 Dynamic Modeling. *See* Simulation

 Ecological Fallacy
 Ecological Validity
 Econometrics
 Effect Size
 Effects Coding
 Effects Coefficient

- Efficiency
 Eigenvalues
 Eigenvector. *See* Eigenvalues; Factor Analysis
 Elaboration (Lazarsfeld's Method of)
 Elasticity
 Emic/Etic Distinction
 Emotions Research
 Empiricism
 Empty Cell
 Encoding/Decoding Model
 Endogenous Variable
 Epistemology
 EQS
 Error
 Error Correction Models
 Essentialism
 Estimation
 Estimator
 Eta
 Ethical Codes
 Ethical Principles
 Ethnograph
 Ethnographic Content Analysis
 Ethnographic Realism
 Ethnographic Tales
 Ethnography
 Ethnomethodology
 Ethnostatistics
 Ethogenics
 Ethology
 Evaluation Apprehension
 Evaluation Research
 Event Count Models
 Event History Analysis
 Event Sampling
 Exogenous Variable
 Expectancy Effect. *See* Experimenter Expectancy Effect
 Expected Frequency
 Expected Value
 Experiment
 Experimental Design
 Experimenter Expectancy Effect
 Expert Systems
 Explained Variance. *See* R-Squared
 Explanation
 Explanatory Variable
 Exploratory Data Analysis
 Exploratory Factor Analysis. *See* Factor Analysis
 External Validity
 Extrapolation
 Extreme Case
F Distribution
F Ratio
 Face Validity
 Factor Analysis
 Factorial Design
 Factorial Survey Method (Rossi's Method)
 Fallacy of Composition
 Fallacy of Objectivism
 Fallibilism
 Falsificationism
 Feminist Epistemology
 Feminist Ethnography
 Feminist Research
 Fiction and Research
 Field Experimentation
 Field Relations
 Field Research
 Fieldnotes
 File Drawer Problem
 Film and Video in Research
 First-Order. *See* Order
 Fishing Expedition
 Fixed Effects Model
 Floor Effect
 Focus Group
 Focused Comparisons. *See* Structured, Focused Comparison
 Focused Interviews
 Forced-Choice Testing
 Forecasting
 Foreshadowing
 Foucauldian Discourse Analysis
 Fraud in Research
 Free Association Interviewing
 Frequency Distribution
 Frequency Polygon
 Friedman Test
 Function
 Fuzzy Set Theory
 Game Theory
 Gamma
 Gauss-Markov Theorem. *See* BLUE; OLS
Geisteswissenschaften
 Gender Issues
 General Linear Models
 Generalizability. *See* External Validity

-
- Generalizability Theory
 - Generalization/Generalizability in Qualitative Research
 - Generalized Additive Models
 - Generalized Estimating Equations
 - Generalized Least Squares
 - Generalized Linear Models
 - Geometric Distribution
 - Gibbs Sampling
 - Gini Coefficient
 - GLIM
 - Going Native
 - Goodness-of-Fit Measures
 - Grade of Membership Models
 - Granger Causality
 - Graphical Modeling
 - Grounded Theory
 - Group Interview
 - Grouped Data
 - Growth Curve Model
 - Guttman Scaling

 - Halo Effect
 - Hawthorne Effect
 - Hazard Rate
 - Hermeneutics
 - Heterogeneity
 - Heteroscedasticity. *See* Heteroskedasticity
 - Heteroskedasticity
 - Heuristic
 - Heuristic Inquiry
 - Hierarchical (Non)Linear Model
 - Hierarchy of Credibility
 - Higher-Order. *See* Order
 - Histogram
 - Historical Methods
 - Holding Constant. *See* Control
 - Homoscedasticity. *See* Homoskedasticity
 - Homoskedasticity
 - Humanism and Humanistic Research
 - Humanistic Coefficient
 - Hypothesis
 - Hypothesis Testing. *See* Significance Testing
 - Hypothetico-Deductive Method

 - Ideal Type
 - Idealism
 - Identification Problem
 - Idiographic/Nomothetic. *See* Nomothetic/Idiographic
 - Impact Assessment
 - Implicit Measures

 - Imputation
 - Independence
 - Independent Variable
 - Independent Variable (in Experimental Research)
 - Independent Variable (in Nonexperimental Research)
 - In-Depth Interview
 - Index
 - Indicator. *See* Index
 - Induction
 - Inequality Measurement
 - Inequality Process
 - Inference. *See* Statistical Inference
 - Inferential Statistics
 - Influential Cases
 - Influential Statistics
 - Informant Interviewing
 - Informed Consent
 - Instrumental Variable
 - Interaction. *See* Interaction Effect;
Statistical Interaction
 - Interaction Effect
 - Intercept
 - Internal Reliability
 - Internal Validity
 - Internet Surveys
 - Interpolation
 - Interpretative Repertoire
 - Interpretive Biography
 - Interpretive Interactionism
 - Interpretivism
 - Interquartile Range
 - Interrater Agreement
 - Interrater Reliability
 - Interrupted Time-Series Design
 - Interval
 - Intervening Variable. *See* Mediating
Variable
 - Intervention Analysis
 - Interview Guide
 - Interview Schedule
 - Interviewer Effects
 - Interviewer Training
 - Interviewing
 - Interviewing in Qualitative Research
 - Intraclass Correlation
 - Intracoder Reliability
 - Investigator Effects
 - In Vivo Coding
 - Isomorph
 - Item Response Theory

- Jackknife Method
- Key Informant
- Kish Grid
- Kolmogorov-Smirnov Test
- Kruskal-Wallis H Test
- Kurtosis
- Laboratory Experiment
- Lag Structure
- Lambda
- Latent Budget Analysis
- Latent Class Analysis
- Latent Markov Model
- Latent Profile Model
- Latent Trait Models
- Latent Variable
- Latin Square
- Law of Averages
- Law of Large Numbers
- Laws in Social Science
- Least Squares
- Least Squares Principle. *See* Regression
- Leaving the Field
- Leslie's Matrix
- Level of Analysis
- Level of Measurement
- Level of Significance
- Levene's Test
- Life Course Research
- Life History Interview. *See* Life Story Interview
- Life History Method
- Life Story Interview
- Life Table
- Likelihood Ratio Test
- Likert Scale
- LIMDEP
- Linear Dependency
- Linear Regression
- Linear Transformation
- Link Function
- LISREL
- Listwise Deletion
- Literature Review
- Lived Experience
- Local Independence
- Local Regression
- Loess. *See* Local Regression
- Logarithm
- Logic Model
- Logical Empiricism. *See* Logical Positivism
- Logical Positivism
- Logistic Regression
- Logit
- Logit Model
- Log-Linear Model
- Longitudinal Research
- Lowess. *See* Local Regression
- Macro
- Mail Questionnaire
- Main Effect
- Management Research
- MANCOVA. *See* Multivariate Analysis of Covariance
- Mann-Whitney U Test
- MANOVA. *See* Multivariate Analysis of Variance
- Marginal Effects
- Marginal Homogeneity
- Marginal Model
- Marginals
- Markov Chain
- Markov Chain Monte Carlo Methods
- Matching
- Matrix
- Matrix Algebra
- Maturation Effect
- Maximum Likelihood Estimation
- MCA. *See* Multiple Classification Analysis
- McNemar Change Test. *See* McNemar's Chi-Square Test
- McNemar's Chi-Square Test
- Mean
- Mean Square Error
- Mean Squares
- Measure. *See* Conceptualization, Operationalization, and Measurement
- Measure of Association
- Measures of Central Tendency
- Median
- Median Test
- Mediating Variable
- Member Validation and Check
- Membership Roles
- Memos, Memoing
- Meta-Analysis
- Meta-Ethnography
- Metaphors
- Method Variance
- Methodological Holism
- Methodological Individualism

-
- Metric Variable
 - Micro
 - Microsimulation
 - Middle-Range Theory
 - Milgram Experiments
 - Missing Data
 - Misspecification
 - Mixed Designs
 - Mixed-Effects Model
 - Mixed-Method Research. *See* Multimethod Research
 - Mixture Model (Finite Mixture Model)
 - MLE. *See* Maximum Likelihood Estimation
 - Mobility Table
 - Mode
 - Model
 - Model I ANOVA
 - Model II ANOVA. *See* Random-Effects Model
 - Model III ANOVA. *See* Mixed-Effects Model
 - Modeling
 - Moderating. *See* Moderating Variable
 - Moderating Variable
 - Moment
 - Moments in Qualitative Research
 - Monotonic
 - Monte Carlo Simulation
 - Mosaic Display
 - Mover-Stayer Models
 - Moving Average
 - Mplus
 - Multicollinearity
 - Multidimensional Scaling (MDS)
 - Multidimensionality
 - Multi-Item Measures
 - Multilevel Analysis
 - Multimethod-Multitrait Research. *See* Multimethod Research
 - Multimethod Research
 - Multinomial Distribution
 - Multinomial Logit
 - Multinomial Probit
 - Multiple Case Study
 - Multiple Classification Analysis (MCA)
 - Multiple Comparisons
 - Multiple Correlation
 - Multiple Correspondence Analysis.
See Correspondence Analysis
 - Multiple Regression Analysis
 - Multiple-Indicator Measures
 - Multiplicative
 - Multistage Sampling
 - Multi-Strategy Research
 - Multivariate
 - Multivariate Analysis
 - Multivariate Analysis of Variance and Covariance (MANOVA and MANCOVA)
 - N (n)
 - N6
 - Narrative Analysis
 - Narrative Interviewing
 - Native Research
 - Natural Experiment
 - Naturalism
 - Naturalistic Inquiry
 - Negative Binomial Distribution
 - Negative Case
 - Nested Design
 - Network Analysis
 - Neural Network
 - Nominal Variable
 - Nomothetic. *See* Nomothetic/Idiographic
 - Nomothetic/Idiographic
 - Nonadditive
 - Nonlinear Dynamics
 - Nonlinearity
 - Nonparametric Random-Effects Model
 - Nonparametric Regression. *See* Local Regression
 - Nonparametric Statistics
 - Nonparticipant Observation
 - Nonprobability Sampling
 - Nonrecursive
 - Nonresponse
 - Nonresponse Bias
 - Nonsampling Error
 - Normal Distribution
 - Normalization
 - NUD*IST
 - Nuisance Parameters
 - Null Hypothesis
 - Numeric Variable. *See* Metric Variable
 - NVivo
 - Objectivism
 - Objectivity
 - Oblique Rotation. *See* Rotations
 - Observation Schedule
 - Observation, Types of
 - Observational Research
 - Observed Frequencies
 - Observer Bias

- Odds
Odds Ratio
Official Statistics
Ogive. *See* Cumulative Frequency Polygon
OLS. *See* Ordinary Least Squares
Omega Squared
Omitted Variable
One-Sided Test. *See* One-Tailed Test
One-Tailed Test
One-Way ANOVA
Online Research Methods
Ontology, Ontological
Open Question. *See* Open-Ended Question
Open-Ended Question
Operational Definition. *See* Conceptualization, Operationalization, and Measurement
Operationism/Operationalism
Optimal Matching
Optimal Scaling
Oral History
Order
Order Effects
Ordinal Interaction
Ordinal Measure
Ordinary Least Squares (OLS)
Organizational Ethnography
Orthogonal Rotation. *See* Rotations
Other, the
Outlier
- p* Value
Pairwise Correlation. *See* Correlation
Pairwise Deletion
Panel
Panel Data Analysis
Paradigm
Parameter
Parameter Estimation
Part Correlation
Partial Correlation
Partial Least Squares Regression
Partial Regression Coefficient
Participant Observation
Participatory Action Research
Participatory Evaluation
Pascal Distribution
Path Analysis
Path Coefficient. *See* Path Analysis
Path Dependence
Path Diagram. *See* Path Analysis
- Pearson's Correlation Coefficient
Pearson's *r*. *See* Pearson's Correlation Coefficient
Percentage Frequency Distribution
Percentile
Period Effects
Periodicity
Permutation Test
Personal Documents. *See* Documents, Types of
Phenomenology
Phi Coefficient. *See* Symmetric Measures
Philosophy of Social Research
Philosophy of Social Science
Photographs in Research
Pie Chart
Piecewise Regression. *See* Spline Regression
Pilot Study
Placebo
Planned Comparisons. *See* Multiple Comparisons
Poetics
Point Estimate
Poisson Distribution
Poisson Regression
Policy-Oriented Research
Politics of Research
Polygon
Polynomial Equation
Polytomous Variable
Pooled Surveys
Population
Population Pyramid
Positivism
Post Hoc Comparison. *See* Multiple Comparisons
Postempiricism
Posterior Distribution
Postmodern Ethnography
Postmodernism
Poststructuralism
Power of a Test
Power Transformations. *See* Polynomial Equation
Pragmatism
Precoding
Predetermined Variable
Prediction
Prediction Equation
Predictor Variable
Pretest
Pretest Sensitization
Priming
Principal Components Analysis
Prior Distribution

-
- Prior Probability
 Prisoner's Dilemma
 Privacy. *See* Ethical Codes; Ethical Principles
 Privacy and Confidentiality
 Probabilistic
 Probabilistic Guttman Scaling
 Probability
 Probability Density Function
 Probability Sampling. *See* Random Sampling
 Probability Value
 Probing
 Probit Analysis
 Projective Techniques
 Proof Procedure
 Propensity Scores
 Proportional Reduction of Error (PRE)
 Protocol
 Proxy Reporting
 Proxy Variable
 Pseudo-*R*-Squared
 Psychoanalytic Methods
 Psychometrics
 Psychophysiological Measures
 Public Opinion Research
 Purposive Sampling
 Pygmalion Effect
- Q Methodology
 Q Sort. *See* Q Methodology
 Quadratic Equation
 Qualitative Content Analysis
 Qualitative Data Management
 Qualitative Evaluation
 Qualitative Meta-Analysis
 Qualitative Research
 Qualitative Variable
 Quantile
 Quantitative and Qualitative Research, Debate About
 Quantitative Research
 Quantitative Variable
 Quartile
 Quasi-Experiment
 Questionnaire
 Queuing Theory
 Quota Sample
- R*
r
 Random Assignment
 Random Digit Dialing
 Random Error
 Random Factor
 Random Number Generator
 Random Number Table
 Random Sampling
 Random Variable
 Random Variation
 Random Walk
 Random-Coefficient Model. *See* Mixed-Effects Model;
 Multilevel Analysis
 Random-Effects Model
 Randomized-Blocks Design
 Randomized Control Trial
 Randomized Response
 Randomness
 Range
 Rank Order
 Rapport
 Rasch Model
 Rate Standardization
 Ratio Scale
 Rationalism
 Raw Data
 RC(M) Model. *See* Association Model
 Reactive Measures. *See* Reactivity
 Reactivity
 Realism
 Reciprocal Relationship
 Recode
 Record-Check Studies
 Recursive
 Reductionism
 Reflexivity
 Regression
 Regression Coefficient
 Regression Diagnostics
 Regression Models for Ordinal Data
 Regression on . . .
 Regression Plane
 Regression Sum of Squares
 Regression Toward the Mean
 Regressor
 Relationship
 Relative Distribution Method
 Relative Frequency
 Relative Frequency Distribution
 Relative Variation (Measure of)
 Relativism
 Reliability
 Reliability and Validity in Qualitative Research

- Reliability Coefficient
Repeated Measures
Repertory Grid Technique
Replication
Replication/Replicability in Qualitative Research
Representation, Crisis of
Representative Sample
Research Design
Research Hypothesis
Research Management
Research Question
Residual
Residual Sum of Squares
Respondent
Respondent Validation. *See* Member Validation and Check
Response Bias
Response Effects
Response Set
Retroduction
Rhetoric
Rho
Robust
Robust Standard Errors
Role Playing
Root Mean Square
Rotated Factor. *See* Factor Analysis
Rotations
Rounding Error
Row Association. *See* Association Model
Row-Column Association. *See* Association Model
R-Squared

Sample
Sampling
Sampling Bias. *See* Sampling Error
Sampling Distribution
Sampling Error
Sampling Fraction
Sampling Frame
Sampling in Qualitative Research
Sampling Variability. *See* Sampling Error
Sampling With (or Without) Replacement
SAS
Saturated Model
Scale
Scaling
Scatterplot
Scheffé's Test
Scree Plot

Secondary Analysis of Qualitative Data
Secondary Analysis of Quantitative Data
Secondary Analysis of Survey Data
Secondary Data
Second-Order. *See* Order
Seemingly Unrelated Regression
Selection Bias
Self-Administered Questionnaire
Self-Completion Questionnaire. *See* Self-Administered Questionnaire
Self-Report Measure
Semantic Differential Scale
Semilogarithmic. *See* Logarithms
Semiotics
Semipartial Correlation
Semistructured Interview
Sensitive Topics, Researching
Sensitizing Concept
Sentence Completion Test
Sequential Analysis
Sequential Sampling
Serial Correlation
Shapiro-Wilk Test
Show Card
Sign Test
Significance Level. *See* Alpha, Significance Level of a Test
Significance Testing
Simple Additive Index. *See* Multi-Item Measures
Simple Correlation (Regression)
Simple Observation
Simple Random Sampling
Simulation
Simultaneous Equations
Skewed
Skewness. *See* Kurtosis
Slope
Smoothing
Snowball Sampling
Social Desirability Bias
Social Relations Model
Sociogram
Sociometry
Soft Systems Analysis
Solomon Four-Group Design
Sorting
Sparse Table
Spatial Regression
Spearman Correlation Coefficient
Spearman-Brown Formula

-
- Specification
 - Spectral Analysis
 - Sphericity Assumption
 - Spline Regression
 - Split-Half Reliability
 - Split-Plot Design
 - S-Plus
 - Spread
 - SPSS
 - Spurious Relationship
 - Stability Coefficient
 - Stable Population Model
 - Standard Deviation
 - Standard Error
 - Standard Error of the Estimate
 - Standard Scores
 - Standardized Regression Coefficients
 - Standardized Test
 - Standardized Variable. *See* z-Score; Standard Scores
 - Standpoint Epistemology
 - Stata
 - Statistic
 - Statistical Comparison
 - Statistical Control
 - Statistical Inference
 - Statistical Interaction
 - Statistical Package
 - Statistical Power
 - Statistical Significance
 - Stem-and-Leaf Display
 - Step Function. *See* Intervention Analysis
 - Stepwise Regression
 - Stochastic
 - Stratified Sample
 - Strength of Association
 - Structural Coefficient
 - Structural Equation Modeling
 - Structural Linguistics
 - Structuralism
 - Structured Interview
 - Structured Observation
 - Structured, Focused Comparison
 - Substantive Significance
 - Sum of Squared Errors
 - Sum of Squares
 - Summated Rating Scale. *See* Likert Scale
 - Suppression Effect
 - Survey
 - Survival Analysis
 - Symbolic Interactionism
 - Symmetric Measures
 - Symmetry
 - Synchronic
 - Systematic Error
 - Systematic Observation. *See* Structured Observation
 - Systematic Review
 - Systematic Sampling
 - Tau. *See* Symmetric Measures
 - Taxonomy
 - Tchebechev's Inequality
 - t*-Distribution
 - Team Research
 - Telephone Survey
 - Testimonio
 - Test-Retest Reliability
 - Tetrachoric Correlation. *See* Symmetric Measures
 - Text
 - Theoretical Sampling
 - Theoretical Saturation
 - Theory
 - Thick Description
 - Third-Order. *See* Order
 - Threshold Effect
 - Thurstone Scaling
 - Time Diary. *See* Time Measurement
 - Time Measurement
 - Time-Series Cross-Section (TSCS) Models
 - Time-Series Data (Analysis/Design)
 - Tobit Analysis
 - Tolerance
 - Total Survey Design
 - Transcription, Transcript
 - Transfer Function. *See* Box-Jenkins Modeling
 - Transformations
 - Transition Rate
 - Treatment
 - Tree Diagram
 - Trend Analysis
 - Triangulation
 - Trimming
 - True Score
 - Truncation. *See* Censoring and Truncation
 - Trustworthiness Criteria
 - t*-Ratio. *See* *t*-Test
 - t*-Statistic. *See* *t*-Distribution
 - t*-Test
 - Twenty Statements Test
 - Two-Sided Test. *See* Two-Tailed Test
 - Two-Stage Least Squares

- Two-Tailed Test
Two-Way ANOVA
Type I Error
Type II Error
- Unbalanced Designs
Unbiased
Uncertainty
Uniform Association. *See* Association Model
Unimodal Distribution
Unit of Analysis
Unit Root
Univariate Analysis
Universe. *See* Population
Unobserved Heterogeneity
Unobtrusive Measures
Unobtrusive Methods
Unstandardized
Unstructured Interview
Utilization-Focused Evaluation
- V.* *See* Symmetric Measures
Validity
Variable
Variable Parameter Models
Variance. *See* Dispersion
Variance Components Models
Variance Inflation Factors
Variation
Variation (Coefficient of)
Variation Ratio
Varimax Rotation. *See* Rotations
Vector
- Vector Autoregression (VAR). *See* Granger
Causality
Venn Diagram
Verbal Protocol Analysis
Verification
Verstehen
Vignette Technique
Virtual Ethnography
Visual Research
Volunteer Subjects
- Wald-Wolfowitz Test
Weighted Least Squares
Weighting
Welch Test
White Noise
Wilcoxon Test
WinMAX
Within-Samples Sum of Squares
Within-Subject Design
Writing
- X* Variable
 $\bar{X}$
- Y*-Intercept
Y Variable
Yates' Correction
Yule's *Q*
- z*-Score
z-Test
Zero-Order. *See* Order

Reader's Guide

ANALYSIS OF VARIANCE

Analysis of Covariance (ANCOVA)
Analysis of Variance (ANOVA)
Main Effect
Model I ANOVA
Model II ANOVA
Model III ANOVA
One-Way ANOVA
Two-Way ANOVA

ASSOCIATION AND CORRELATION

Association
Association Model
Asymmetric Measures
Biserial Correlation
Canonical Correlation Analysis
Correlation
Correspondence Analysis
Intraclass Correlation
Multiple Correlation
Part Correlation
Partial Correlation
Pearson's Correlation
 Coefficient
Semipartial Correlation
Simple Correlation (Regression)
Spearman Correlation Coefficient
Strength of Association
Symmetric Measures

BASIC QUALITATIVE RESEARCH

Autobiography
Life History Method
Life Story Interview
Qualitative Content Analysis
Qualitative Data Management
Qualitative Research

Quantitative and Qualitative Research,
 Debate About
Secondary Analysis of Qualitative Data

BASIC STATISTICS

Alternative Hypothesis
Average
Bar Graph
Bell-Shaped Curve
Bimodal
Case
Causal Modeling
Cell
Covariance
Cumulative Frequency Polygon
Data
Dependent Variable
Dispersion
Exploratory Data Analysis
F Ratio
Frequency Distribution
Histogram
Hypothesis
Independent Variable
Median
Measures of Central Tendency
N(*n*)
Null Hypothesis
Pie Chart
Regression
Standard Deviation
Statistic
t-Test
 $\bar{X}$
Y Variable
z-Test

CAUSAL MODELING

Causality
Dependent Variable

Effects Coefficient
Endogenous Variable
Exogenous Variable
Independent Variable
Path Analysis
Structural Equation Modeling

DISCOURSE/CONVERSATION ANALYSIS

Accounts
Conversation Analysis
Critical Discourse Analysis
Deviant Case Analysis
Discourse Analysis
Foucauldian Discourse Analysis
Interpretative Repertoire
Proof Procedure

ECONOMETRICS

ARIMA
Cointegration
Durbin-Watson Statistic
Econometrics
Fixed Effects Model
Mixed-Effects Model
Panel
Panel Data Analysis
Random-Effects Model
Selection Bias
Serial Correlation (Regression)
Time-Series Cross-Section (TSCS) Models
Time-Series Data (Analysis/Design)
Tobit Analysis

EPISTEMOLOGY

Constructionism, Social
Epistemology
Idealism
Interpretivism
Laws in Social Science
Logical Positivism
Methodological Holism
Naturalism
Objectivism
Positivism

ETHNOGRAPHY

Autoethnography
Case Study
Creative Analytical Practice (CAP) Ethnography

Critical Ethnography
Ethnographic Content Analysis
Ethnographic Realism
Ethnographic Tales
Ethnography
Participant Observation

EVALUATION

Applied Qualitative Research
Applied Research
Evaluation Research
Experiment
Heuristic Inquiry
Impact Assessment
Qualitative Evaluation
Randomized Control Trial

EVENT HISTORY ANALYSIS

Censoring and Truncation
Event History Analysis
Hazard Rate
Survival Analysis
Transition Rate

EXPERIMENTAL DESIGN

Experiment
Experimenter Expectancy Effect
External Validity
Field Experimentation
Hawthorne Effect
Internal Validity
Laboratory Experiment
Milgram Experiments
Quasi-Experiment

FACTOR ANALYSIS AND RELATED TECHNIQUES

Cluster Analysis
Commonality Analysis
Confirmatory Factor Analysis
Correspondence Analysis
Eigenvalue
Exploratory Factor Analysis
Factor Analysis
Oblique Rotation
Principal Components Analysis
Rotated Factor

Rotations
Varimax Rotation

FEMINIST METHODOLOGY

Feminist Ethnography
Feminist Research
Gender Issues
Standpoint Epistemology

GENERALIZED LINEAR MODELS

General Linear Models
Generalized Linear Models
Link Function
Logistic Regression
Logit
Logit Model
Poisson Regression
Probit Analysis

HISTORICAL/COMPARATIVE

Comparative Method
Comparative Research
Documents, Types of
Emic/Etic Distinction
Historical Methods
Oral History

INTERVIEWING IN QUALITATIVE RESEARCH

Biographic Narrative Interpretive
Method (BNIM)
Dependent Interviewing
Informant Interviewing
Interviewing in Qualitative Research
Narrative Interviewing
Semistructured Interview
Unstructured Interview

LATENT VARIABLE MODEL

Confirmatory Factor Analysis
Item Response Theory
Factor Analysis
Latent Budget Analysis
Latent Class Analysis
Latent Markov Model
Latent Profile Model

Latent Trait Models
Latent Variable
Local Independence
Nonparametric Random-Effects Model
Structural Equation Modeling

LIFE HISTORY/BIOGRAPHY

Autobiography
Biographic Narrative Interpretive
Method (BNIM)
Interpretive Biography
Life History Method
Life Story Interview
Narrative Analysis
Psychoanalytic Methods

LOG-LINEAR MODELS (CATEGORICAL DEPENDENT VARIABLES)

Association Model
Categorical Data Analysis
Contingency Table
Expected Frequency
Goodness-of-Fit Measures
Log-Linear Model
Marginal Model
Marginals
Mobility Table
Odds Ratio
Saturated Model
Sparse Table

LONGITUDINAL ANALYSIS

Cohort Analysis
Longitudinal Research
Panel
Period Effects
Time-Series Data (Analysis/Design)

MATHEMATICS AND FORMAL MODELS

Algorithm
Assumptions
Basic Research
Catastrophe Theory
Chaos Theory
Distribution
Fuzzy Set Theory
Game Theory

MEASUREMENT LEVEL

Attribute
Binary
Categorical
Continuous Variable
Dichotomous Variables
Discrete
Interval
Level of Measurement
Metric Variable
Nominal Variable
Ordinal Measure

**MEASUREMENT TESTING
AND CLASSIFICATION**

Conceptualization, Operationalization,
and Measurement
Generalizability Theory
Item Response Theory
Likert Scale
Multiple-Indicator Measures
Summated Rating Scale

MULTILEVEL ANALYSIS

Contextual Effects
Dependent Observations
Fixed Effects Model
Mixed-Effects Model
Multilevel Analysis
Nonparametric Random-Effects
Model
Random-Coefficient Model
Random-Effects Model

MULTIPLE REGRESSION

Adjusted *R*-Squared
Best Linear Unbiased Estimator
Beta
Generalized Least Squares
Heteroskedasticity
Interaction Effect
Misspecification
Multicollinearity
Multiple Regression Analysis
Nonadditive
R-Squared
Regression
Regression Diagnostics

Specification
Standard Error of the Estimate

QUALITATIVE DATA ANALYSIS

Analytic Induction
CAQDAS
Constant Comparison
Grounded Theory
In Vivo Coding
Memos, Memoing
Negative Case
Qualitative Content Analysis

SAMPLING IN QUALITATIVE RESEARCH

Purposive Sampling
Sampling in Qualitative Research
Snowball Sampling
Theoretical Sampling

SAMPLING IN SURVEYS

Multistage Sampling
Quota Sampling
Random Sampling
Representative Sample
Sampling
Sampling Error
Stratified Sampling
Systematic Sampling

SCALING

Attitude Measurement
Bipolar Scale
Dimension
Dual Scaling
Guttman Scaling
Index
Likert Scale
Multidimensional Scaling (MDS)
Optimal Scaling
Scale
Scaling
Semantic Differential Scale
Thurstone Scaling

SIGNIFICANCE TESTING

Alpha, Significance Level of a Test
Confidence Interval

Level of Significance
One-Tailed Test
Power of a Test
Significance Level
Significance Testing
Statistical Power
Statistical Significance
Substantive Significance
Two-Tailed Test

SIMPLE REGRESSION

Coefficient of Determination
Constant
Intercept
Least Squares
Linear Regression
Ordinary Least Squares
Regression on . . .
Regression
Scatterplot
Slope
Y-Intercept

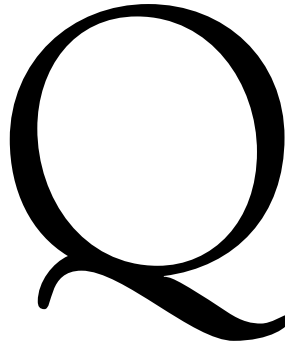
SURVEY DESIGN

Computer-Assisted Personal
Interviewing
Internet Surveys

Interviewing
Mail Questionnaire
Secondary Analysis
of Survey Data
Structured Interview
Survey
Telephone Survey

TIME SERIES

ARIMA
Box-Jenkins Modeling
Cointegration
Detrending
Durbin-Watson Statistic
Error Correction Models
Forecasting
Granger Causality
Interrupted Time-Series Design
Intervention Analysis
Lag Structure
Moving Average
Periodicity
Serial Correlation
Spectral Analysis
Time-Series Cross-Section (TSCS)
Models
Time-Series Data (Analysis/Design)
Trend Analysis



Q METHODOLOGY

Q methodology provides a framework for a science of subjectivity that incorporates procedures for data collection (Q-sort technique) and analysis (FACTOR ANALYSIS). Introduced by William Stephenson, Q methodology typically involves collecting a UNIVERSE of opinions on some topic (e.g., environmental activism) from which a REPRESENTATIVE Q sample of 30 to 50 statements is selected, such as, “No one, however virtuous the cause, should be above the law,” “I am a bit suspicious of their motives,” and so on. Two or three dozen participants then sort the items from agree (+5) to disagree (−5), and the Q sorts are CORRELATED and factor analyzed using the PQMethod or PCQ programs (accessible at www.qmethod.org), thereby revealing the diversity of subjectivities at issue. Factor scores associated with the statements provide the basis for interpreting the factors.

In a study of environmental activism, for example, statements such as the above were administered to three dozen members of the British public, whose Q sorts revealed seven distinct narratives, such as law-abidingness, liberal humanism, radical activism, and so on (Capdevila & Stainton Rogers, 2000). In a single-case study of a dissociative disorder, each of several members of a multiple personality provided Q sort perceptions of relationships to other members of the system, the factorization of which revealed the organization of the personality (Smith, 2000, pp. 336–338).

Q methodology parallels quantum mechanics in conceptual and mathematical respects, as summarized in a series of five articles by Stephenson (*Psychological Record*, 1986–1988). Much subjective behavior in literature, politics, decision making, psychotherapy, newspaper reading, and all other areas of human endeavor is PROBABILISTIC, indeterminate, and transitive, but takes definite form when subjected to Q sorting, in which meaning and measurement are inseparable in the Q sorter’s acts of judgment. The number of factors is uncertain a priori, and they are in a relationship of complementarity.

Although developed within BEHAVIORISM, the wide applicability of Q methodology has led to its adoption by postmodernists, social CONSTRUCTIONISTS, feminists, DISCOURSE and narrative analysts, cognitive scientists, psychoanalysts, geographers, and both quantitative and qualitative researchers. It has also been taken up by policy analysts because of its facility in revealing stakeholder perspectives (Brown, Durning, & Selden, 1999), and by those interested in emergent democratic identities (Dryzek & Holmes, 2002). Continuing scholarship appears in the pages of *Operant Subjectivity*; the journal of the International Society for the Scientific Study of Subjectivity; in the *Journal of Human Subjectivity*; and in *Q-Methodology and Theory*, the Korean-language journal of the Korean Society for the Scientific Study of Subjectivity.

—Steven R. Brown

REFERENCES

- Brown, S. R., Durning, D. W., & Selden, S. C. (1999). Q methodology. In G. J. Miller & M. L. Whicker (Eds.), *Handbook of research methods in public administration* (pp. 599–637). New York: Dekker.
- Capdevila, R., & Stainton Rogers, R. (2000). If you go down to the woods today...: Narratives of Newbury. In H. Addams & J. Proops (Eds.), *Social discourse and environmental policy: An application of Q methodology* (pp. 152–173). Cheltenham, UK: Elgar.
- Dryzek, J. S., & Holmes, L. T. (2002). *Post-communist democratization*. Cambridge, UK: Cambridge University Press.
- Smith, N. W. (2000). *Current systems in psychology*. Belmont, CA: Wadsworth.

Q SORT. See Q METHODOLOGY

QUADRATIC EQUATION

The *quadratic equation* is an elementary algebraic method for solving equations where the variable of interest is squared. This is a useful technique for obtaining “roots” (the point where the squared variable of interest crosses the x -axis).

More formally, a quadratic equation refers to second-degree algebraic equations. It may have one or more variables, but the standard form in a single variable x is given by

$$ax^2 + bx + c = 0, \quad (1)$$

where a , b , and c are constant in the equation (referred to as coefficients). The equation is simplified when the coefficients are equal to zero, but the quadratic equation remains of the same form provided $a \neq 0$. For example:

$$x^2 - 3x + 2 = 0 \quad \text{and} \quad 2x^2 - x - 3 = 0 \quad (2)$$

are typical quadratic equations. The solution to a quadratic equation is found by completing the square according to

$$x^2 + \frac{b}{a}x = -\frac{c}{a}$$

$$x + \frac{b^2}{2a} = -\frac{c}{a} + \frac{b^2}{4a^2} = \frac{b^2 - 4ac}{4a^2},$$

$$x + \frac{b}{2a} = \frac{\pm\sqrt{b^2 - 4ac}}{2a},$$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \quad (3)$$

Equation (3) is called a quadratic formula, which helps us solve the quadratic equation of interest.

Returning to Equation (2), the quadratic equation is parsed as follows:

$$2x^2 - x - 3 = 0 \rightarrow a = 2, \quad b = -1, \quad c = -3.$$

This leads to the solution

$$x = \frac{-(-1) \pm \sqrt{(-1)^2 - [4 \times 2 \times (-3)]}}{2 \times 2}$$

$$= \frac{1 \pm 5}{4} = \frac{3}{2} \quad \text{or} \quad -1.$$

In this construction, the subquantity $b^2 - 4ac$ is called discriminant and is usually designated with the Greek letter δ . The discriminant of a quadratic equation determines the nature of its solutions (also called roots). For instance, if

$$b^2 - 4ac > 0,$$

the quadratic equation has two distinct real roots:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Furthermore, if

$$b^2 - 4ac = 0,$$

the quadratic equation has only one real root:

$$x = -\frac{b}{2a}.$$

If, on the other hand,

$$b^2 - 4ac < 0,$$

then there are no real roots for x . The solution lies in the complex numbers.

The roots of quadratic equations can also be visualized from the graph of its corresponding quadratic function. Consider

$$y = ax^2 + bx + c,$$

where again a , b , and c are constant and $a \neq 0$. The x versus y graph of a quadratic function is called a

parabola, which is bowl-like and symmetric about a vertical line. The roots of the original quadratic equation ($ax^2 + bx + c = 0$) are the values of x where the parabola intersects with the x -axis. A quadratic function may have one (the original quadratic equation has two identical roots), two (the original quadratic equation has two distinctive real roots), or no x -intercepts (the original quadratic equation has no real roots), which is ultimately determined by the discriminant, as described above.

A more general form of the quadratic equation is

$$ax^2 + by^2 + cxy + dx + ey + f = 0, \quad (4)$$

where a, b, c, d, e , and f are constants and at least one of a, b , or c is nonzero. For instance,

$$4x^2 + y^2 - 4xy - 2x + y + 1 = 0$$

is a general quadratic equation for x and y . The graph of this type of quadratic equation is called a conic section. It can be a hyperbola, a parabola, an ellipse, intersecting lines, a point, distinctive parallel lines, or coincident lines. The following method is used to determine which type of conic section a general quadratic equation generates (Zwillinger, 1995):

$$\Delta = \begin{vmatrix} a & \frac{c}{2} & \frac{d}{2} \\ \frac{c}{2} & b & \frac{e}{2} \\ \frac{d}{2} & \frac{e}{2} & f \end{vmatrix} \quad J = \begin{vmatrix} a & \frac{c}{2} \\ \frac{c}{2} & b \end{vmatrix},$$

$$I = a + b, \quad K = \begin{vmatrix} a & \frac{d}{2} \\ \frac{d}{2} & f \end{vmatrix} + \begin{vmatrix} b & \frac{c}{2} \\ \frac{c}{2} & f \end{vmatrix}.$$

The typology in Table 1 maps parameter values to the described geometric forms.

Quadratic equations can be even more general with additional variables, but the resulting multidimensional graphical forms can produce quite complicated surfaces.

—Zhen Li

REFERENCE

Zwillinger, D. (1995). *CRC standard mathematical tables and formulae* (30th ed.). Boca Raton, FL: CRC Press.

Table 1

Δ	J	Δ/I	K	Conic Form
$\neq 0$	< 0			Hyperbola
$\neq 0$	0			Parabola
$\neq 0$	> 0	< 0		Ellipse
0	< 0			Intersecting lines
0	> 0			Point
0	0		< 0	Distinct parallel lines
0	0		0	Coincident lines

QUALITATIVE CONTENT ANALYSIS

Qualitative content analysis examines significant aspects of texts that are not amenable to quantitative techniques. Such techniques measure patterns of frequency and regularity in a large number of texts, but as Siegfried Kracauer (1952–1953) once argued in a key paper, what is perhaps most significant about a particular TEXT may resist quantification. Even where this is not the case, some aspects of a text cannot be counted easily, and when they can, this may tell us little of how they operate within or across texts. Examples of textual features and functions resistant, if not allergic, to quantification include irony, ambivalence, and allusion; communicative register and mode of address; folkloric motifs, aesthetic codes, and generic conventions; rhetorical and stylistic devices, including resonant METAPHORS and other figures of speech; and the point of view, presuppositions, and values that may come implicitly with the message and make certain categories or notions appear natural or absolute in meaning.

To take a specific case, certain presuppositions and values may coalesce in shifting a particular category onto the terrain of stereotyping and constructions of the OTHER. Analyzing the construction and operation of a stereotypical figure or conception requires the formulation of a theoretically informed argument as well as the combination of different methodological procedures (see Pickering, 2001).

Qualitative content analysis begins at the point where statistical presentation reaches its limits, and does so in order to reveal the significance of

textual features that are latent or hidden in the manifest content or that have consequences beyond their immediate, obtrusive meaning. Even a single lexical choice, such as a headline in a newspaper or a loaded term in a government report, may reverberate throughout the whole text in a series of interactions with other components of the text, and a particular text may set off a chain of development and response that escapes its reference in a quantified pattern of frequency. For Kracauer, such qualities of a text create its texture of intimations and implications.

Two of the most important qualitative approaches to the study of media content and cultural texts have been SEMIOTICS and CRITICAL DISCOURSE ANALYSIS (see Deacon, Pickering, Golding, & Murdock, 1999, Chaps. 7 and 8). Both approaches offer various conceptual tools that can be used in unpicking textures of intimation and implication in a document or artifact, showing how they have broader ramifications within the cultural formations in which they are produced and disseminated. Both also show how particular significations and representations are multiaccented or intersected by different currents of meaning and value. Such forms of qualitative content analysis do not treat cultural texts as isolated entities but as sites of interaction between processes of production and processes of interpretation. As such, they bear the definite imprint of their production but are never fixed in their meaning. They provide cues for their interpretation but are open to multiple readings. Qualitative modes of analysis are best for tracking the perspectives from which cultural texts start and the directions to which they lead as they are constructed and reconstructed, and so come to mean what they always provisionally mean, even as they may seem to insist on a fixed, or singular, legitimate interpretation.

—Michael J. Pickering

See also CONTENT ANALYSIS

REFERENCES

- Deacon, D., Pickering, M., Golding, P., & Murdock, G. (1999). *Researching communications*. London: Arnold.
- Kracauer, S. (1952–1953). The challenge of qualitative content analysis. *Public Opinion Quarterly*, 16, 631–642.
- Pickering, M. (2001). *Stereotyping: The politics of representation*. Basingstoke, UK: Palgrave.

QUALITATIVE DATA MANAGEMENT

Data management is a challenging, integral, and vital part of qualitative research if it is to be successful. Good management has been identified as necessary for facilitating the coherence of a project (Huberman & Miles, 1994). Managing data well “facilitates interpretation just as good orchestration facilitates good dance music” (Meadows & Dodendorf, 1999, p. 196). Researchers need a plan from the outset to sort, summarize, analyze, and store project data, including their process of working with the data through the iterative process. The iterative progress of qualitative research means that data management and data analysis are integral to each other.

Data management in qualitative research begins with the conceptualization of the project. Although what constitutes data per se varies with some established traditions, the process of doing qualitative research, and doing it well, means that the researcher’s reflexivity is part of the project data. Formulating the research question from a topic of interest based on reflection and investigations of existing literature (and potentially theory) are central to the project. A thorough literature review is the first step in the qualitative analysis (Morse & Richards, 2002). The research question guides the data to be collected: Without good data, one cannot have a good study. In the study design stage, decisions made regarding sampling, recruitment, data collection techniques, and analytic approaches are steps in data management that need to be recorded and enveloped into the project data.

Data usually come from several sources and can be copious, including documents, pictures, media clips, and transcripts as well as field notes and reflective journals at a very early stage. These need to be documented, turned into derived text or other forms that can be analyzed, and used as data early and often. Researchers must remember to back up all data being used in the project. Descriptions of codes, categories or themes, and memos that reflect the process of moving data from description to analysis and the final interpretation and/or theory-building process are all part of the data in a project.

Qualitative analytic programs are an integral part of this process (Meadows & Dodendorf, 1999), facilitating the management of projects large or small in a way that allows the researcher to spend time on analysis and

interpretation instead of searching for important quotes or that elusive file with important codes. Data can be managed manually in file drawers, diagrams, charts, models, and tables that can become cumbersome, preventing the research team from closely working with its data. Programs can also aid the process of asking questions of the coded data to develop theory or aid the interpretive process.

Finally, management also includes adhering to ethical standards that include anonymizing data, ensuring that only named members of research teams have access to the data, and that data in all their forms are kept secure throughout the project. Careful planning for data management at an early stage in the research process can make a significant contribution to the rigor and success of qualitative research.

—Lynn M. Meadows

See also DATA MANAGEMENT

REFERENCES

- Huberman, M. A., & Miles, M. B. (1994). Data management & analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 428–444). Thousand Oaks, CA: Sage.
- Meadows, L. M., & Dodendorf, D. M. (1999). Data management and interpretation: Using computers to assist. In B. F. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (pp. 195–218). Thousand Oaks, CA: Sage.
- Morse, J. M., & Richards, L. (2002). *Read me first for a user's guide to qualitative methods*. Thousand Oaks, CA: Sage.

QUALITATIVE EVALUATION

In judging the effectiveness of a program, evaluators often conduct open-ended interviews, observe programs, review documents, and construct case studies. Such methods constitute qualitative evaluation.

Qualitative evaluations emphasize reporting program participants' experiences in *their own words*. Evaluators often gather statistical data, but it is also important to understand the stories and perspectives of the real people who make up the numbers. Qualitative evaluations also go beyond assessing goal attainment to capture unintended impacts and ripple effects, and to illuminate dimensions of desired outcomes that are difficult to quantify, such as what it *means* to a participant to have achieved a goal.

Qualitative methods are especially useful in evaluating programs that individualize outcomes, which means matching services to the varying specific needs of individual participants. In such programs, outcomes can involve qualitatively different dimensions for different clients. Qualitative evaluations can do individualized case studies to document outcome variations.

Process evaluations often use qualitative methods to study *how* a program produces outcomes. Process evaluations aim at elucidating and understanding the internal dynamics of how a program operates by investigating the following kinds of questions: What things do people experience that make this program what it is? How are clients brought into the program, and how do they move through the program once they are participants? How is what people do related to what they are trying to (or actually do) accomplish? What are the strengths and weaknesses of the program from the perspective of participants and staff?

Qualitative evaluation is highly appropriate for studying program processes because (a) depicting process requires detailed descriptions of how people engage with each other; (b) participants' experiences of process typically vary, so their different perspectives need to be captured; and (c) process is fluid and dynamic, so it cannot be fairly summarized on a single rating scale at one point in time. Process evaluations examine both formal activities and informal or unplanned interactions.

Qualitative methods are also used for implementation evaluations to document how a program has developed and how and why programs may deviate from initial plans and expectations. Other evaluation concerns for which qualitative methods are especially appropriate include comparing diverse programs, documenting development over time, investigating system changes, and personalizing and humanizing evaluation (Patton, 2002).

Participatory evaluations often rely heavily on qualitative methods because they are easier to understand than statistics. In participatory evaluations, the evaluator works collaboratively with program staff and/or participants to assess program effectiveness. Open-ended interviewing is a common method in participatory evaluations because those involved can learn to gather and analyze the data with the evaluator's help and guidance. Involving participants collaboratively can increase the usefulness of evaluations (Patton, 1997).

Qualitative methods can be used by themselves or in combination with quantitative techniques in mixed-methods designs. Qualitative methods, although once controversial, have become widely used and valued in evaluation studies.

—Michael Quinn Patton

See also EVALUATION RESEARCH

REFERENCES

- Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.
- Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks, CA: Sage.

QUALITATIVE META-ANALYSIS

Also referred to as *qualitative metasynthesis* (Sandelowski, Docherty, & Emden, 1997), *qualitative meta-data-analysis* (Paterson, Thorne, Canam, & Jillings, 2001), and *META-ETHNOGRAPHY* (Noblit & Hare, 1988), qualitative meta-analysis is a distinctive category of synthesis in which the findings from completed qualitative studies in a target area are formally combined. Both an analytic process and an interpretive product, qualitative meta-analysis is analogous to quantitative META-ANALYSIS in its intent to ascertain systematically, comprehensively, and transparently the state of knowledge in a field of study. Owing in part to the early stage of its development as a research method, qualitative meta-analysis has been variously portrayed as a form of, encompassed by, or synonymous with other kinds of qualitative synthesis, such as GROUNDED THEORY, metastudy, and SECONDARY ANALYSIS OF QUALITATIVE DATA.

Key factors accounting for the increasing interest in conducting qualitative meta-analysis studies include the proliferation of qualitative studies in the health and social sciences and practice disciplines, a related concern that the important findings from these studies are underutilized in practice and policy forums, and a desire to counter the depiction of evidence and evidence-based practice as deriving solely from experimental research. Emphasizing IDIOGRAPHIC over NOMOTHETIC knowledge, or case-bound over formal generalization, qualitative meta-analysis is a means toward enhancing the relevance and utility of qualitative research because it entails the intensive,

case-oriented study of phenomena in larger and more varied samples than are typically the rule in any one qualitative study. Yet the goal of qualitative meta-analysis is to preserve the particulars of experience represented in individual studies. Especially in practice and policy forums, knowledge of the particular is seen to be critical to offset the frequent failure of nomothetic generalizations to fit the individual case.

But it is the very emphasis of qualitative research on the particular that has created one of the greatest challenges to integrating research results. Also complicating the effort to sum up the findings of qualitative studies is the diversity of research approaches and agendas encompassed by what is commonly referred to as QUALITATIVE RESEARCH. Other challenges in conducting qualitative meta-analyses include defining the substantive limits of study; retrieving relevant studies; determining whether, or the extent to which, such features as method influence the findings; judging the quality of a study; determining whether a study constitutes qualitative research; and, given the diversity in styles of reporting qualitative research, finding the results in qualitative studies. There remains also the question of which analytic, interpretive, and representational techniques are best suited to synthesizing the findings from individual studies and reporting the results of a meta-analysis study to diverse audiences.

Despite these challenges, efforts toward moving qualitative research into the research integration scene are growing. Most notable among these are the increasing publication of qualitative meta-analysis studies, funding of research to develop qualitative meta-analysis techniques, and progress toward inclusion of qualitative research in Campbell and Cochrane databases of reviews of research.

—Margarete Sandelowski

See also META-ANALYSIS

REFERENCES

- Noblit, G. W., & Hare, R. D. (1988). *Meta-ethnography: Synthesizing qualitative studies*. Newbury Park, CA: Sage.
- Paterson, B. L., Thorne, S. E., Canam, C., & Jillings, C. (2001). *Meta-study of qualitative health research: A practical guide to meta-analysis and meta-synthesis*. Thousand Oaks, CA: Sage.
- Sandelowski, M., Docherty, S., & Emden, C. (1997). Qualitative metasynthesis: Issues and techniques. *Research in Nursing & Health*, 20, 365–371.

QUALITATIVE RESEARCH

Qualitative research is an umbrella term for an array of attitudes toward and strategies for conducting inquiry that are aimed at discerning how human beings understand, experience, interpret, and produce the social world (Mason, 1996). Although typically juxtaposed with QUANTITATIVE RESEARCH, qualitative research is not a unified form of inquiry clearly differentiated from it, but rather home to a variety of scholars from the sciences, humanities, and practice disciplines committed to different and, sometimes, conflicting philosophical and methodological positions.

These philosophical positions include, but are not limited to, INTERPRETIVISM, HERMENEUTICS, SOCIAL CONSTRUCTIONISM, and CRITICAL THEORY, as these are variously defined. Methodologic approaches include, but are not limited to, PHENOMENOLOGY, GROUNDED THEORY, ETHNOGRAPHY, participatory inquiry, and NARRATIVE and DISCOURSE ANALYSIS, as these are variously defined (Denzin & Lincoln, 2000a). History, literary criticism, ethics, and cultural and material culture studies are examples of disciplines often viewed as wholly defined by qualitative research. Yet demographic history, for example, is largely quantitative, and scholars in these fields do not typically refer to themselves as qualitative researchers or to their work as qualitative research, or even as research at all. Qualitative researchers in the social sciences have turned to the arts and humanities to learn how to read and perform texts like artists and humanists, whereas qualitative researchers in the humanities have turned to the social sciences to learn how to theorize and conduct fieldwork like social scientists (Denzin & Lincoln, 2000b). In short, *qualitative research* is a term that tends to obscure both difference and commonality. Qualitative research cannot be understood and, indeed, is often misunderstood and misrepresented by simply contrasting it with quantitative research, which is, in turn, similarly misrepresented by efforts to reduce its diversity and complexity.

QUALITATIVE RESEARCH VERSUS QUALITATIVE DATA AND TECHNIQUES

What is collectively referred to as qualitative research should be distinguished from other research and other works that merely contain qualitative data,

such as descriptive SURVEY studies and journalistic accounts, and from data collection and analysis techniques commonly viewed as qualitative research techniques, such as FOCUS GROUPS and CONTENT ANALYSIS. Both focus groups and content analysis have their origins in, and are still widely used in, quantitative research. That is, there is no qualitative attitude or strategy necessarily inherent in these techniques. Moreover, all quantitative behavioral and social science research with human subjects depends on qualitative data in the QUESTIONNAIRES employed. The mere fact that words are used or produced in a project, or that words prevail over numbers, does not make a work qualitative research, just as the mere fact that numbers are used or produced in a project does not make a work quantitative research.

DEFINING FEATURES OF QUALITATIVE RESEARCH

What makes a work deserving of the label *qualitative research* is the demonstrable effort to produce richly and relevantly detailed descriptions and particularized interpretations of people and the social, linguistic, material, and other practices and events that shape and are shaped by them. Qualitative research typically includes, but is not limited to, discerning the perspectives of these people, or what is often referred to as the actor's point of view. Although both philosophically and methodologically a highly diverse entity, qualitative research is marked by certain defining imperatives that include its case (as opposed to its variable) orientation, sensitivity to cultural and historical context, and REFLEXIVITY. In its many guises, qualitative research is a form of empirical inquiry that typically entails some form of PURPOSIVE SAMPLING for information-rich cases; IN-DEPTH INTERVIEWS and open-ended interviews, lengthy participant/field observations, and/or document or artifact study; and techniques for analysis and interpretation of data that move beyond the DATA generated and their surface appearances. Qualitative research is typically directed toward producing IDIOGRAPHIC (as opposed to NOMOTHETIC) knowledge because its emphasis is on the penetrating understanding of particular phenomena, events, or cases.

In contrast to what are collectively referred to as quantitative researchers, who aim for a disciplined OBJECTIVITY in their attitude toward subjects and their collection and treatment of data, qualitative researchers

aim for a disciplined subjectivity that embraces even as it makes explicit the partiality inherent in all inquiry. The emphasis in what is collectively referred to as quantitative research is on the control of research conditions to minimize bias and threats to the validity of findings, and commitment to a RESEARCH DESIGN developed prior to entering the field of study. In contrast, the emphasis in what is collectively referred to as qualitative research is on NATURALISM—that is, observing events as they unfold without manipulating any conditions—and the ongoing development and refinement of the research design after entering the field of study. As Mason (1996) insightfully noted, qualitative research requires people who are able to “think and act strategically,” flexibly, and creatively while in the field to “combine intellectual, philosophical, technical, practical (and ethical) concerns” (p. 2). Qualitative research is also characterized by adherence to a diverse array of orientations to and strategies for maximizing the VALIDITY or TRUSTWORTHINESS of study procedures and results, as well as the use of expressive language and more expressly literary representational styles for disseminating findings.

MANY APPROACHES, ONE FEELING TONE

Although what is collectively encompassed in the term *qualitative research* resists simple definition, it still manages to convey one “feeling tone” (Sapir, 1951, p. 308) of reverence for individuality, diversity, and everyday life, and the hope that inquiry can bridge the gaps that divide people. Qualitative research thus designates not only a domain of inquiry but also a site of protest and reconciliation.

—Margarete Sandelowski

REFERENCES

- Denzin, N. K., & Lincoln, Y. S. (Eds.). (2000a). *Handbook of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.
- Denzin, N. K., & Lincoln, Y. S. (2000b). The discipline and practice of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 1–28). Thousand Oaks, CA: Sage.
- Mason, J. (1996). *Qualitative researching*. London: Sage.
- Sapir, E. (1951). Culture, genuine and spurious. In D. G. Mandelbaum (Ed.), *Selected writings of Edward Sapir in language, culture and personality* (pp. 308–331). Berkeley: University of California Press.

QUALITATIVE VARIABLE

A variable whose value varies by attributes or characteristics is called a qualitative variable. The scale for measurement of a qualitative variable is a set of *unordered* or *nominal* categories. For example, gender (female, male); party identification (Democratic, Republican, Independent); marital status (single, married, divorced, widowed); ethnic group (with categories such as white, black, Hispanic, Asian, other); employment status (unemployed, employed, retired); and religious affiliation (with categories such as Catholic, Jewish, Protestant, Muslim, Buddhist, other, none) are all qualitative variables.

Qualitative variables differ from quantitative variables in that distinct categories matter in quality or attribute rather than in quantity or magnitude. In other words, no category is assumed to be greater than or smaller than any of the other categories. For example, “male” and “female” are merely labels for one’s gender. These two labels do not indicate any difference in quantity or magnitude. Likewise, “Methodist,” “Catholic,” “Presbyterian,” “Muslim,” and “Buddhist” are labels for identifying one’s religious affiliation. None of them represents greater or smaller magnitudes of the variable “religious affiliation.”

In social research, one can use qualitative variables to classify data with respect to certain characteristics and make decisions about which elements are most similar and most different. Thus, the level of measurement for qualitative variables is on a nominal scale. The term “NOMINAL” implies that there are no numerical differences between categories. If a variable can vary in numerical values among subjects in a SAMPLE or POPULATION, then this variable is called a *quantitative* variable rather than a qualitative variable. QUANTITATIVE VARIABLES are measured on an *interval* scale because there is a specific numerical distance, or interval, between each pair of levels. For example, a respondent’s age can be measured on an interval scale. The interval between an age of 30 and an age of 50 equals 20. Thus, we can compare the ages of two people to determine how much older or how much younger one is than the other. Such comparisons are impossible and do not make sense to qualitative variables.

It is useful to understand a third type of measurement scale that lies in between interval and nominal. If the values of a variable have a natural *ordering*

for an *ordinal* scale but undefined interval distances between any pair of categories, then that variable is called an ordinal variable. For example, one's political ideology can be measured as very liberal, slightly liberal, moderate, slightly conservative, or very conservative. This collection of ordered categories has a clear ordering, but the distances between categories are undefined. For instance, although we know that a person self-identified as "very liberal" is more liberal than a person categorized as "conservative," we cannot know the numerical value for how much more liberal that person is.

Ordinal variables and qualitative variables are usually called *categorical variables*. However, statistical methods for analyzing ordinal and qualitative variables are usually different. Data analysts need to know the interval-ordinal-nominal scale classification and quantitative-qualitative classification because each statistical method refers to a particular type of data. For example, summarizing data using the MEAN and STANDARD DEVIATION of a variable is applicable for quantitative variables but is inappropriate for analyzing qualitative variables. Notice that researchers usually assign discrete numbers to the categories or labels of a qualitative variable for purposes of CODING (for example, white = 1, black = 2, Hispanic = 3, Asian = 4, others = 5). However, the numbers assigned to particular classifications of data cannot be manipulated in the ways one analyzes quantitative and ordinal variables. More information on statistical methods for CATEGORICAL variables may be found in Alan Agresti (2002), J. Scott Long (1997), and Daniel Powers and Yu Xie (2000).

—Cheng-Lung Wang

REFERENCES

- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley Interscience.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Powers, D., & Xie, Y. (2000). *Statistical methods for categorical data analysis*. San Diego, CA: Academic Press.

QUANTILE

A quantile is the proportion (or percentage) of DATA points below the given value. For example,

0.25 (or 25%) is the point at which 25% of the observations fall below and 75% fall above that value. Quantile has many useful applications, from a quantile-quantile (q-q) plot to quantile REGRESSION.

—Tim Futing Liao

QUANTITATIVE AND QUALITATIVE RESEARCH, DEBATE ABOUT

The debate about quantitative and qualitative research is concerned with the question of whether the two research strategies should be considered contrasting epistemological positions or whether they are better conceptualized as simply referring to different clusters of techniques of data collection and analysis. The issue is very significant in terms of the prospects for MULTISTRATEGY RESEARCH. If quantitative and qualitative research are viewed as distinctive and largely irreconcilable epistemologies, multistrategy research is very difficult to envision. On the other hand, if they are viewed as different methods of data collection and analysis, research combining quantitative and qualitative research (i.e., multistrategy research) is easier to imagine.

The epistemological version of the debate has two major forms (Bryman, 2001). One is the *embedded methods* argument. This argument asserts that any research method is rooted in epistemological commitments. Thus, to choose to conduct research using a SELF-ADMINISTERED QUESTIONNAIRE is not just seeking answers to a research question in a certain way; the choice also entails a commitment to the epistemological (and ontological) principles with which it is deemed to be associated (for example, POSITIVISM). When Hughes (1990) writes that "every research tool or procedure is inextricably embedded in commitments to particular visions of the world and to knowing that world" (p. 11), he was essentially taking up the embedded methods argument.

The other form of the epistemological version of the debate is the *paradigm* argument. Drawing on Kuhn's (1970) notion of the PARADIGM, this argument depicts quantitative and qualitative research as paradigms. As such, epistemological assumptions, values, and methods are viewed as inseparable within each of the two paradigms. In addition, because paradigms are deemed to be incapable of reconciliation, quantitative

and qualitative research cannot be reconciled. This form of the epistemological version shares with the embedded methods argument the view that methods of research are linked to epistemological positions, but it goes further in viewing quantitative and qualitative research more generally as incommensurable.

Although the epistemological version of the debate about quantitative and qualitative research still has adherents who rail against the notion that the two research strategies can be combined at anything more than the most superficial level (e.g., Sale, Lohfeld, & Brazil, 2002), this position has increasingly given way to the technical version of the debate. Writers associated with the technical version recognize that quantitative and qualitative research have connections with epistemological and ontological assumptions but do not see these connections as inevitable. In other words, research methods can serve different masters, so that a “research method from one research strategy is viewed as capable of being pressed into the service of another” (Bryman, 2001, p. 446).

This softening among most writers in their views about quantitative and qualitative research, whereby the technical version of the debate is tending to hold sway, has almost certainly led to an increase in multistrategy research and certainly in its acceptability.

—Alan Bryman

REFERENCES

- Bryman, A. (2001). *Social research methods*. Oxford, UK: Oxford University Press.
- Hughes, J. A. (1990). *The philosophy of social research* (2nd ed.). Harlow, UK: Longman.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.
- Sale, J. E. M., Lohfeld, L. H., & Brazil, K. (2002). Revisiting the quantitative-qualitative debate: Implications for mixed-methods research. *Quality and Quantity*, 36, 43–53.

QUANTITATIVE RESEARCH

Research can be grouped into two different types: quantitative and qualitative. The difference has to do with the ways in which the data are collected and how many observations there are. Sample surveys are examples of quantitative research, whereas small ethnographic case studies are examples of qualitative research.

A quantitative research project is characterized by having a population for which the researcher wants to draw conclusions, but it is not possible to collect data on the entire population. For an observational study, it is necessary to select a proper, statistical RANDOM SAMPLE and to use methods of STATISTICAL INFERENCE to draw conclusions about the population. For an experimental study, it is necessary to have a RANDOM ASSIGNMENT of subjects to experimental and control groups in order to use methods of statistical inference.

Statistical methods are used in all three stages of a quantitative research project. For observational studies, the data are collected using statistical sampling theory. Then, the sample data are analyzed using descriptive statistical analysis. Finally, generalizations are made from the sample data to the entire population using statistical inference. For experimental studies, the subjects are allocated to experimental and control group using randomizing methods. Then, the experimental data are analyzed using descriptive statistical analysis. Finally, just as for observational data, generalizations are made to a larger population.

HISTORICAL ACCOUNT

The previous century saw large growth in the area of theory and methods for the drawing of statistically random samples. The British statistician Sir Ronald A. Fisher was an early proponent for the use of randomness in the collection of data, particularly in experiments. In experimental research, there has also been large growth in the area of designs for the running of experiments. Each design carries with it its own computational methods, and they are more accessible in both statistics text and statistical software.

The second half of the previous century saw the arrival of large, national sample studies. The studies done by George Gallup and his associates are the well-known examples of such studies, but there are also other early studies of this kind. The Roper studies are an example. The best-known academic studies have been done by the National Opinion Research Center (NORC), first located at the University of Chicago, and by the Survey Research Center at the Institute for Social Research at the University of Michigan. NORC has collected data for many years in their General Social Survey, and the Michigan group has a long history of the study of economic behavior and voting in elections.

APPLICATIONS

Drawing a national sample of respondents is a difficult task because there is no accessible list of the residents of this country. Various sophisticated methods of telephone sampling are now commonly used, but because of their complexity, special statistical formulas have been developed for the analysis of such data. The formulas in all introductory statistics texts and most statistical computer packages are based on SIMPLE RANDOM SAMPLING of respondents. But the STANDARD ERRORS computed from simple random sampling are not correct when the sample has been collected using more complicated methods, as they often are for large, national studies.

EXAMPLES

After the data have been collected, either through SAMPLING or experiments, statistics is used to analyze the data. This is typically done by arranging the data in tables, making graphs of the data, or computing summary statistics for the data. Often, two or all three methods are used. These are methods used by quantitative research as well as qualitative research. In quantitative research, the data often have to be weighed in various ways to compensate for the ways in which the data were collected.

Herein lies the largest difference between quantitative research and qualitative research. Because the data are collected randomly, only quantitative research can make use of methods of statistical inference. It is not as clear how generalizations are made on the basis of qualitative research. Results can be used to obtain anecdotal evidence, but neither point estimates nor CONFIDENCE INTERVALS and *P* VALUES can be used. However, insights gathered through qualitative research often go deeper than what is possible with quantitative research. They can also act as the foundation for future quantitative research.

In the computation of both *p* values and confidence intervals, it is important to use formulas that reflect the way in which the data were collected. In experimental research, it is common to compute a variety of sums of squares from the data and use these sums of squares for the computation of *F* values and thereby *p* values. Similar formulas are not as easily available for more complicated sample surveys.

As implied by its name, quantitative research results in the computation of a variety of quantities, most often of a numerical nature. With proper random data, it

is possible to compute summaries such as MEANS and MEDIANS. They, in turn, help us understand the data on the corresponding variables. The same holds for *p* values and conclusions from statistical inference.

Critics point out that such summary quantities can be too restrictive and may not tell the story the way it should be told. The many examples of statistical fallacies and statistical jokes are reflections of this. Even in quantitative research, it becomes necessary to consider special cases, such as OUTLIERS, to gain a better understanding of the data. As soon as we move from statistical summaries to individual observations, we really move from quantitative research to the richness of qualitative research. Such studies of individual data points blur the line between the two approaches. One particular area where the shortcomings of quantitative research can show up is in REGRESSION studies of the relationships between variables. A single data point that is away from the cluster of the remaining observations can have a large influence on both the regression line and the correlation coefficient. On the other hand, a quantity such as the median is not as sensitive to outliers.

—Gudmund R. Iversen

REFERENCES

- Bickman, L., & Rog, D. J. (1998). *Handbook of applied social research methods*. Thousand Oaks, CA: Sage.
- Brown, S. R., & Melamed, L. E. (1998). *Experimental design and analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-074). Thousand Oaks, CA: Sage.
- Kalton, G. (1983). *Introduction to survey sampling* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-035). Beverly Hills, CA: Sage.

QUANTITATIVE VARIABLE

Variables used in statistical analyses can be listed in increased complexity. Quantitative variables are the most complex because they are measured using a unit of measurement. For any two observations on a quantitative variable, they have either the same or different values, as is the case for a NOMINAL VARIABLE. Also, the values can be ranked from low to high, as is the case for an ordinal variable. In addition, because a quantitative variable has a unit of measurement, two

observations can be different, and it is possible to measure *how* different the observations are. Quantitative variables can be either INTERVAL or ratio variables. More and stronger statistical methods are available for the analysis of quantitative variables than for nominal and ordinal variables, REGRESSION analysis being the most important such method.

Examples of quantitative variables are age, income, and number of children in a household. The values of such variables are measured with the help of a unit of measurement. The unit commonly used for age is a year. That way, one person can be 32 years old, and another person can be 25 years old and therefore 7 years younger than the other person. The unit for income is often the dollar, but 1,000 is also used. Number of children is a counting variable, and the common unit is one child.

However, there is nothing about a variable such as age that necessarily makes it a quantitative variable. The complexity of this variable will depend on the way in which it is measured. It could be measured using the values *young*, *middle aged*, and *old*. Measured this way, age is *not* a quantitative variable; it is an ordinal variable. Also, it could be measured using the values *middle aged* and *not middle aged*. That way it would have to be considered a nominal variable. Thus, in addition to the name of the variable, it is equally necessary to know the way in which the variable is measured.

A less important feature of a quantitative variable is that it can be either DISCRETE or CONTINUOUS. A discrete variable is one for which it is possible to find two values of the variable such that there are no values between the two. Counting variables are discrete; no household can have between two and three children. Similarly, a continuous variable is such that between any two values, it is always possible to find another value. Age, for example, is a continuous variable as we move through time, even though it is often measured as the age at the previous birthday. Sometimes, quantitative variables are called continuous variables.

Quantitative variables are better than nominal or ordinal variables because they can be used for computations such as addition and multiplication. That opens the possibilities for a large variety of statistical computations and analyses. It is possible to find the mean value of a quantitative variable, and it is possible to use the variables for bivariate and multivariate analyses.

—Gudmund R. Iversen

REFERENCES

- JMP Introductory Guide*. (2000). Cary, NC: SAS Institute.
 Moore, D. S., & McCabe, G. P. (1998). *Introduction to the practice of statistics*. New York: W. H. Freeman.
SPSS base 10.0 user's guide. (1999). Chicago: SPSS.

QUARTILE

OBSERVATIONS can be grouped into four equal-sized sets according to their RANK ORDER. Each of the four sets forms a quartile, which is a special case of QUANTILE.

—Tim Futing Liao

See also QUANTILE

QUASI-EXPERIMENT

Quasi-experiments manipulate presumed causes to discover their effects, but the researcher does *not* assign units to conditions randomly. Quasi-experiments are necessary because it is not always possible to randomize. Ethical constraints may preclude withholding effective treatments from needy people based on chance without proper informed consent, those who administer treatment may refuse to honor randomization, or questions about program effects may arise after a TREATMENT was already implemented so that randomization is impossible. So, quasi-experiments use a combination of design features, practical logic, and statistical analysis to show that the treatment may be a plausible cause of the effect. The resulting causal inferences are often more ambiguous than is the case with randomized experiments. Nonrandomized experiment is synonymous with quasi-experiment, and observational study and nonexperimental design often include quasi-experiments as a subset.

KINDS OF QUASI-EXPERIMENTAL DESIGNS

Quasi-experimental designs include, but are not limited to, (a) nonequivalent control group designs, in which the outcomes of those exposed to two or more conditions are studied but the experimenter does not control assignment to conditions; (b) INTERRUPTED TIME-SERIES DESIGNS, in which many consecutive observations over time (prototypically 100) are

available on an outcome, and treatment is introduced in the midst of those observations to demonstrate its impact on the outcome through a discontinuity in the time series after treatment; (c) regression discontinuity designs, in which the experimenter uses a cutoff score on a measured variable to determine eligibility for treatment, and an effect is observed if the regression line of the assignment variable on outcome for the treatment group is discontinuous from that of the comparison group; and (d) single-case designs, in which one participant is observed repeatedly over time (usually on fewer occasions than in the time series) while the scheduling and dose of treatment are manipulated to demonstrate that treatment controls outcome.

In such designs, treatment is manipulated, and outcome is then observed. Two other classes of designs are sometimes included as quasi-experiments, even though the cause is not manipulated. In (e) case control designs, a group with an outcome of interest is compared to a group without that outcome to see if they differ retrospectively in exposure to possible causes; and in (f) correlational designs, observations on possible treatments and outcomes are observed simultaneously, often with a survey, to see if they are related. Because these designs do not ensure that cause precedes effect, as it must logically do, they usually yield more equivocal causal inferences.

HISTORICAL DEVELOPMENT

Most experiments conducted prior to the 1920s were quasi-experiments. For example, Lind (1753) described a quasi-experimental comparison of six medical treatments for scurvy; around 1850, epidemiologists used case control methods to identify contaminated water supplies as the cause of cholera in London; and in 1898, Triplett used a nonequivalent control group design to show that the presence of an audience and competitors improved the performance of bicyclists.

In 1963, Campbell and Stanley coined the term *quasi-experiment* to describe this class of designs. Campbell and his colleagues (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002) extended the theory and practice of these designs in three ways. First, they described a larger number of these designs. For example, some quasi-experimental designs are inherently longitudinal (e.g., time series, single-case designs), observing participants over time; but other designs can be made longitudinal by adding more

observations before or after treatment. Similarly, more than one treatment or control group can be used, and the designs can be combined, as when adding a nonequivalent control group to a time series.

Second, Campbell developed a logic to evaluate the quality of causal inferences resulting from quasi-experimental designs—a VALIDITY typology elaborated in Cook and Campbell (1979) and Shadish et al. (2002). The typology includes four validity types and threats to validity for each type. Threats are common reasons why researchers may be wrong about the causal inferences they draw. Statistical conclusion validity concerns inferences about how much presumed cause and effect covary; an example of a threat is low STATISTICAL POWER. INTERNAL VALIDITY concerns inferences that observed covariation is due to the treatment causing the outcome; a sample threat is history (extraneous events that could also cause the effect). CONSTRUCT VALIDITY concerns inferences about higher-order constructs that research operations represent; a sample threat is EXPERIMENTER EXPECTANCY EFFECTS, whereby participants react to what they believe the experimenter wants to observe rather than to the intended treatment. EXTERNAL VALIDITY concerns inferences about whether the cause-effect relationship holds over variation in people, settings, treatment variables, and measurement variables; threats include interactions of the treatment with other features of the design that produce unique effects that would not be observed otherwise.

Third, Campbell addressed threats to validity using design features—things a researcher can manipulate to prevent a threat from occurring or to diagnose its presence and impact on study results (Table 1). For example, suppose maturation (normal development over time) is an anticipated threat to validity because it could cause a pretest-posttest change like that attributed to treatment. The inclusion of several consecutive pretests before treatment can indicate whether the rate of maturation before treatment is similar to the rate of change during and after treatment. If so, maturation is a threat. All quasi-experiments are combinations of these design features, chosen to diagnose or rule out threats to validity in a particular context. Campbell was skeptical about adjusting threats statistically after they have already occurred because statistical adjustments require making assumptions that are usually dubious or impossible to test.

Other scholars during this time were also interested in causal inferences in quasi-experiments, particularly

Table 1 Design Elements Used in Constructing Quasi-Experiments**Assignment (Control of assignment strategies to increase group comparability)**

- *Cutoff-Based Assignment.* Controlled assignment to conditions based on one or more fully measured covariates. This yields an unbiased effect estimate.
- *Other Nonrandom Assignment.* Various forms of “haphazard” assignment that sometimes approximate randomization (e.g., alternating assignment in a two-condition quasi-experiment whereby every other unit is assigned to one condition, etc.).
- *Matching and Stratifying.* Efforts to create groups equivalent on observed covariates in ways that are stable, do not lead to regression artifacts, and are correlated with the outcome.

Measurement (Use of measures to learn whether threats to causal inference actually operate)*Posttest Observations*

- *Nonequivalent Dependent Variables.* Measures that are not sensitive to the causal forces of the treatment, but *are* sensitive to all or most of the confounding causal forces that might lead to false conclusions about treatment effects (if such measures show no effect, but the outcome measures do show an effect, the causal inference is bolstered because it is less likely to be due to the confounds).
- *Multiple Substantive Posttests.* Used to assess whether the treatment affects a complex pattern of theoretically predicted outcomes.

Pretest Observations

- *Single Pretest.* A pretreatment measure on the outcome variable, useful to help diagnose selection bias.
- *Retrospective Pretest.* Reconstructed pretests when actual pretests are not feasible—by itself, a very weak design feature, but sometimes better than nothing.
- *Proxy Pretest.* When a true pretest is not feasible, a pretest on a variable correlated with the outcome—often weak by itself.
- *Multiple Pretest Time Points on the Outcome.* Helps reveal pretreatment trends or regression artifacts that might complicate causal inference.
- *Pretests on Independent Samples.* When a pretest is not feasible on the treated sample, one is obtained from a randomly equivalent sample.
- *Complex Predictions Such as Predicted Interaction.* Successfully predicted interactions lend support to causal inference because alternative explanations become less plausible.
- *Measurement of Threats to Internal Validity.* Help diagnose the presence of specific threats to the inference that A caused B, such as whether units actively sought out additional treatments outside the experiment.

Comparison Groups [Selecting comparisons that are “less nonequivalent” or that bracket the treatment group at the pretest(s)]

- *Single Nonequivalent Groups.* Compared to studies without control groups, using a nonequivalent control group helps identify many plausible threats to validity.
- *Multiple Nonequivalent Groups.* Serve several functions. For instance, groups are selected that are as similar as possible to the treated group, but at least one outperforms it initially and at least one underperforms it, thus bracketing the treated group.
- *Cohorts.* Comparison groups chosen from the same institution in a different cycle (e.g., sibling controls in families or last year’s students in schools).
- *Internal (vs. External) Controls.* Plausibly chosen from within the same population (e.g., within the same school rather than from a different school).

Treatment (Manipulations of the treatment to demonstrate that treatment variability affects outcome variability)

- *Removed Treatments.* Shows that an effect diminishes if treatment is removed.
- *Repeated Treatments.* Reintroduces treatments after they have been removed from some group—common in laboratory sciences or where treatments have short-term effects.
- *Switching Replications.* Reverses treatment and control group roles so that one group is the control while the other receives treatment, but the controls receive treatment later while the original treatment group receives no further treatment or has treatment removed.
- *Reversed Treatments.* Provides a conceptually similar treatment that reverses an effect (e.g., reducing access to a computer for some students but increasing access for others).
- *Dosage Variation.* Demonstrates that outcome responds systematically to different levels of treatment.

William G. Cochran in statistics, James J. Heckman in economics, and Sir Austin Bradford Hill in epidemiology. However, Campbell's work was unique for its extensive emphasis on design rather than statistical analysis, its theory of how to evaluate causal inferences, and its sustained development of quasi-experimental theory and method over four decades.

EXAMPLES

In 1966, the Canadian province of Ontario initiated a formal program for the screening and treatment of infants born with phenylketonuria (PKU) to prevent retardation. After the start of the program, 44 infants born with PKU experienced no retardation, and three did. Of these, two were missed by the screening program (Webb et al., 1973). Statistics from prior years suggested a much higher rate of retardation attributable to PKU. Although this study lacked a control group, the authors concluded that the program successfully treated PKU infants. Based on such results, this program was widely adopted in Canada and the United States. This study was a pretest-posttest quasi-experiment with no control group.

In July 1982, Arizona implemented legislation mandating severe penalties for driving while intoxicated; a comparison of monthly results from January 1976 to July 1982 (the control condition) with monthly totals between July 1982 and May 1984 (the treatment condition) found a decrease in traffic fatalities after the new law was implemented. A similar finding occurred in monthly data trends in San Diego after January 1982, when that city implemented a California law that also penalized intoxicated drivers. In a control time series in El Paso, Texas, a city that had no relevant change in its driving laws during this period, monthly fatality trends showed no comparable changes during the months of either January or July 1982. The changes in trends over time in both San Diego and Arizona, compared to the absence of similar trends in El Paso, suggest that the new laws reduced fatalities (West, Hepworth, McCall, & Reich, 1989). This study was an interrupted time-series quasi-experiment.

STATISTICS AND QUASI-EXPERIMENTAL DESIGN

Statisticians such as Paul Holland, Paul Rosenbaum, and Donald Rubin emphasize the need to measure what would have happened to treatment participants

without treatment (the counterfactual) and focus on statistics that can improve estimates of the counterfactual without randomization. A central method uses PROPENSITY SCORES, a predicted probability of group membership obtained from LOGISTIC REGRESSION of actual group membership on predictors of outcome or of how participants got into treatment. MATCHING, stratifying, or covarying on the propensity score can balance nonequivalent groups on those predictors, but those methods cannot balance groups for unobserved variables, so hidden bias may remain. Hence, these statisticians have developed sensitivity analyses to measure how much hidden bias would be necessary to change an effect in important ways.

Economists such as James Heckman and his colleagues have pursued another development, SELECTION BIAS modeling, which aims to remove hidden bias from quasi-experiments by modeling the selection process. Unfortunately, these models have not been very successful in matching results from randomized experiments. Most recently, economists have improved results by combining selection bias models with propensity scores. This topic continues to develop.

A third development is the use of STRUCTURAL EQUATION MODELING (SEM) to study causal relationships in quasi-experiments; this effort has also been only partly successful (Bollen, 1989). The capacity of SEM to model latent variables can sometimes reduce bias caused by unreliability of measurement, but its capacity to generate unbiased effect estimates is hampered by the same lack of knowledge of selection that thwarts selection bias models. A related but newer literature on causality is promising, using directed graphs to help understand the issues (Pearl, 2000).

CONCLUSION

Quasi-experiments are rarely able to provide the confidence about causal inference that randomized experiments can provide, and overreliance on quasi-experiments in some areas is a serious problem. Nevertheless, three important factors have created conditions to improve quasi-experiments. First, extensive practical experience with quasi-experimental designs has provided a database from which we can conduct empirical studies of the theory (Shadish, 2000). Second, after decades of focus on randomized designs, statisticians and economists have turned their attention to improving quasi-experimental designs. Third,

the computer revolution provided both theorists and practitioners with increased capacity to invent and use more sophisticated and computationally intense methods for improving quasi-experiments.

—William R. Shadish and M. H. Clark

REFERENCES

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.
- Cochran, W. G. (1965). The planning of observational studies in human populations. *Journal of the Royal Statistical Society, Series A*, 128, 134–155.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- Hill, A. B. (1953). Observation and experiment. *New England Journal of Medicine*, 248, 995–1001.
- Lind, J. (1753). *A treatise of the scurvy: Of three parts containing an inquiry into the nature, causes and cure of that disease*. Edinburgh, UK: Sands, Murray and Cochran.
- Pearl, J. (2000). *Causality, models, reasoning and inference*. New York: Cambridge University Press.
- Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer-Verlag.
- Schlesselman, J. J. (1982). *Case-control studies: Design, conduct, analysis*. New York: Oxford University Press.
- Shadish, W. R. (2000). The empirical program of quasi-experimentation. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy* (pp. 13–35). Thousand Oaks, CA: Sage.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
- Tripllett, N. (1898). The dynamogenic factors in pacemaking and competition. *American Journal of Psychology*, 9, 507–533.
- Webb, J. F., Khazen, R. S., Hanley, W. B., Partington, M. S., Percy, W. J. L., & Rathborn, J. C. (1973). PKU screening—is it worth it? *Canadian Medical Association Journal*, 108, 328–329.
- West, S. G., Hepworth, J. T., McCall, M. A., & Reich, J. W. (1989). An evaluation of Arizona's July 1982 drunk driving law: Effects on the City of Phoenix. *Journal of Applied Social Psychology*, 19, 1212–1237.
- Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 659–707.

QUESTIONNAIRE

All scientists use “tools” to measure the phenomena of interest to their disciplines, and social scientists are no exception. Thus, just as a telescope serves astronomers and an electron microscope serves microbiologists, the questionnaire serves many social scientists as their primary measurement tool. The challenge to these researchers is to calibrate a questionnaire that will gather data as accurately and efficiently as possible.

From a “total error” perspective, a poorly constructed questionnaire can contribute BIAS and VARIANCE to the data that are gathered (cf. Groves, 1989). In particular, poorly worded, poorly ordered, and/or poorly formatted questions can lead to significant measurement error in the form of both bias and variance, and/or can lead to significant item NONRESPONSE error in the form of bias.

One of the defining characteristics of a questionnaire is whether it is for a SELF-ADMINISTERED survey, as in a mail survey or an INTERNET SURVEY, or whether it is to be *interviewer-administered*, as in a telephone survey or in-person survey. If the questionnaire is self-administered, then the “end user” is the respondent. In these cases, all instructions required to complete the questionnaire accurately must be stated clearly and explicitly within the questionnaire itself. Or, in the case of computerized self-administered questionnaires, instructions can be presented via audio recordings. There are myriad other design features that are affected by whether the questionnaire is self-administered or interviewer-administered; for detailed information, see Dillman (2000).

In creating a particular questionnaire *item*, there are four structural factors to consider: (a) the *question stem* wording; (b) whether the *response option(s)* will be open-end or closed-end; and, if closed-end, whether they are (c) forced-choice and/or (d) balanced. In wording the question stem (i.e., the interrogative being asked), use as few and as simple words possible to convey the meaning of the construct being measured. In operationalizing the construct through this wording, the researcher also must keep in mind the educational level of the respondents. Whenever a questionnaire is self-administered, respondents' literacy must be considered.

Open-end questions allow a respondent to answer in her or his own words and, thereby, provide richer, more detailed responses about what is being measured, especially when administered by an interviewer who can probe for fuller answers in an unbiased and nondirective manner (cf. Fowler & Mangione, 1990). However, open-end responses must be coded before they can be systematically analyzed, and this can be a labor-intensive and costly enterprise to do reliably.

Closed-end responses typically are of a multiple-choice format in which the respondent is provided a series of *exhaustive* and *mutually exclusive* choices from which to choose one. When administered by an interviewer, some closed-end questions have implied response choices, such as “Yes/No” questions. Multiple answers are allowed for some types of questions (e.g., a question that asks about someone’s three favorite leisure time activities). For those interested in the cognitive processes a respondent uses in making sense of a survey question and then deciding how to answer it, see Tourangeau, Rips, and Rasinski (2000).

In designing response options for most closed-end items, the researcher must consider whether the set of choices should contain a “neutral midpoint” and/or an “uncertain” (i.e., “Don’t Know”) response category. Response sets that do not allow the respondent to use a neutral midpoint (e.g., the middle category in the following response set: Agree, *Neither Agree nor Disagree*, Disagree) or do not let a respondent indicate that he or she is “Uncertain” are termed *forced-choice* items. Unforced-choice items are those that provide respondents with an explicit “out,” such as simply indicating he or she is “Undecided.” Choosing which approach to take in wording a survey item should be linked to (a) the phenomenon being measured and (b) respondents’ abilities and willingness to finely differentiate answers (see Krosnick, 1999, for a discussion of “satisficing”).

Another structural consideration in designing a question is whether or not the response scale is balanced or unbalanced. A balanced set of responses has a true midpoint, regardless of whether the midpoint itself is offered as one of the choice. Thus, the response choices of *Strongly Agree*, *Agree*, *Disagree*, or *Strongly Disagree* are balanced even though a true midpoint is not offered explicitly. In contrast, a response set such as *Excellent*, *Good*, *Fair*, or *Poor* is unbalanced because two of the choices are positive, one is neutral, and one is negative.

The order in which questions or blocks of questions are placed within the questionnaire can affect the data that are generated (see Schuman & Presser, 1981). Generally, it is best to organize a questionnaire so that it begins with easy-to-answer, rapport-building items, even if these items will not be analyzed. This reduces the chances of respondents starting the questionnaire, but then quickly refusing to complete it. Next would come items measuring opinions and attitudes, followed by those measuring behaviors and knowledge. The reasoning for this is that answers to behavioral and knowledge questions are thought to be less affected by the order than are opinion and attitudinal questions. The questionnaire should finish with personal background and demographic items. This flow of content within a questionnaire is reasoned to be optimal in minimizing nonresponse, while simultaneously making it less likely that any one type of item will affect answers to subsequent questions.

It is highly recommended that all questionnaires be pilot tested before being used for gathering the actual study data. Pilot testing provides important feedback on how well respondents understand the wording of the questions, on the length of the questionnaire, and on respondent burden. In many cases, it is wise to engage in a round of cognitive interviewing, with one-on-one “think aloud” sessions with pilot respondents who articulate the thinking processes they go through while trying to comprehend and answer each question.

Finally, in devising a questionnaire, the researcher should have a priori plans for data reduction and how to handle MISSING DATA. Data reduction includes the process of using data from several items to create new variables, such as SCALE scores. Missing data are handled in several ways, including statistically imputing a value for each missing datum.

—Paul J. Lavrakas

REFERENCES

- Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method*. New York: Wiley.
- Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized survey interviewing*. Newbury Park, CA: Sage.
- Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
- Krosnick, J. A. (1999). Survey methodology. *Annual Review of Psychology*, 50, 537–567.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

QUEUEING THEORY

Social scientists often study situations in which there is excess demand for some set of limited resources. Queueing theory is a subfield of STOCHASTIC modeling and is the process by which demanders of a specific service or resource are assumed to wait for accommodation. The methodology is widely used in many literatures to describe assembly line-type processes and services in which the completion time is indeterminant. Developed initially to analyze telephone traffic and industrial throughput (Brockmeyer, Halstrom, & Jensen, 1960; Erlang, 1935), it interprets the results from imposing a set of rules that describes three aspects of provision: the manner in which demanders arrive in the system, the capabilities and resources of the supplier, and how individual cases are treated relative to others.

The first component of the model stipulates that transactions arrive stochastically according to an assumed parametric form. Queueing theory models are much easier to describe if we can claim that these arrivals are independent and serial (not arriving simultaneously). It is not required that these transactions stay in the assigned or selected queue until served; they can leave immediately due to queue length (balking), leave after waiting a specified time without being served (reneging), or possibly change to another available queue in the system (jockeying). The parameterized arrivals rate can change with the time index (nonstationarity), or remain fixed over time (stationarity). In addition, different characteristics can be attached to transactions that denote things like priority, complexity, or flags indicating behavior once in the system.

The second major component of the model is a description of the system's servicing resources. Attributes include the number of service queues (or channels), the existence of priority queues, layers of service requirements, and buffering or storage capacity. These features and the way they are arranged determine how efficiently the system can process incoming transactions.

The third and final part of the basic structure is the set of rules governing queue discipline. These include the order in which queued transactions are served; priority arrangements; and, possibly, allowances for "cheating." The two most common arrangements for transaction service are first-in/first-out and first-in/last-out. In addition, priority arrangements can be defined focus on the handling of important transactions as they enter the system; preemptive-priority cases get served directly, usually by displacing a transaction being served, whereas non-preemptive-priority cases are placed at the front of the queue and therefore are accommodated by the next available server.

There exists specialized notation that describes the assumptions of any given queueing theory model called Kendall notation (named after Sir David Kendall). This has the form $A/B/\sigma/k/M/Z$, where A is the distribution of arrivals, B is the service time distribution, σ is the number of servers, k is the total number of transactions that can be either served or stored in the queue, M is the population size from which arrivals are drawn, and Z is a description of the queue discipline.

So, for example, $M/D/3/8/40,000/FIFO$ might describe a barber shop with exponential arrivals, deterministic service (identical cuts), 3 barbers, 5 seats for waiting customers, in a medium-sized town, and serving in order of arrival. It is typical, however, to drop the last three terms in the list when they are obvious from the context of the problem, especially the population size, which cannot be controlled and often does not affect the model specification.

The most straightforward and common model specification is $M/M/\sigma$: exponentially distributed interarrival time (therefore, Poisson distributed arrivals for time period t), exponentially distributed service times, and σ servers. Most texts (cf. Gross & Harris, 1998) start with these systems because, with mild assumptions, many common empirical situations can be described effectively.

—Jeff Gill

REFERENCES

- Brockmeyer, E., Halstrom, H. L., & Jensen, A. (1960). *The life and works of A. K. Erlang*. Kobenhavn: Akademiet for de Tekniske Videnskaber.
- Erlang, A. K. (1935). *Firccifrede logaritmetavler og andre regnetavler til brug ved undervisning og i praksis*. Kobenhavn: G. E. C. Gads.

Gross, D., & Harris, C. M. (1998). *Fundamentals of queueing theory* (3rd ed.). New York: Wiley.

QUOTA SAMPLE

Quota sampling is a method of selecting respondents for surveys. Survey administrators assign quotas to interviewers that define groups of respondents by using a few key demographic characteristics. Interviewers fill their quotas by choosing individuals whose characteristics match these characteristics as respondents. When the responses are combined for all of the respondents, the characteristics of the sample precisely match specified demographic characteristics of the survey population. For example, an interviewer may be assigned a quota of four African Americans, eight whites, two Hispanics, and two Asian Americans, half of which live in the central city and half of which live in the suburbs. When all of the interviews are completed, the aggregate statistics for the SAMPLE and the study POPULATION will match exactly for the demographic characteristics that were included as interview criteria. However, matching on aggregate statistics does not ensure that the sample will accurately reflect the population on the variables that are of interest in the study.

The most infamous example of the BIAS that can occur with quota sampling involves polling for the 1948 presidential election. This election pitted Harry Truman, the Democratic candidate who had assumed the presidency when Franklin D. Roosevelt died, against Thomas Dewey, the Republican candidate. Three polling firms—Roper, Gallup, and Crossley—all declared Dewey as the winner based on interviews that had been conducted using quota sampling methods. Of course, Truman won with just under 50% of the popular vote, beating Dewey by nearly five percentage points. The quota samples were carefully matched to census data on age, race, gender, rent paid, and residence, but the interviewers had chosen too many Republicans. Republicans were easier to interview, either more willing to participate or more approachable, and thus Democrats were underrepresented in the interviews. After that election, newspapers were filled with stories about the failure of the pollsters, as well as many commentators. Subsequently, all three polling firms dropped quota sampling techniques.

The failure to accurately predict the 1948 election provided an impetus toward the use of PROBABILITY SAMPLES and very limited use of quota samples for social science research as well as polling. Most pre-election polling relies on RANDOM DIGIT DIALING (RDD) to select households from which adult respondents are selected at random, or random computer-generated selection of a sample from an automated list of eligible voters. Specific individuals are identified by these survey procedures, and interviewers are not allowed to substitute respondents, thereby eliminating interviewer discretion and selection bias.

BIAS AND QUOTA SAMPLING

The principal reason that the quota samples may not provide accurate information about the population is that the interviewers have discretion over who is interviewed. Whenever human judgment is allowed to determine who is selected, even though, in this case, the interviewers must meet demographic quotas, unintentional bias creeps into the selection process. Although the bias is unintentional, it can cause significant errors in the study data and the estimates (e.g., means, proportions, or REGRESSION COEFFICIENTS) that are produced from the data.

Quota samples are NONPROBABILITY SAMPLES, which are sometimes referred to as PURPOSEFUL or CONVENIENCE SAMPLING. Nonprobability samples have the common characteristic that human judgment has a role in determining which individuals are selected to respond to the interviewers' questions. Nonprobability samples are often contrasted with PROBABILITY SAMPLES, where each individual in the study population has a (known), nonzero chance of being selected as a RESPONDENT. Probability samples rely on a random mechanism to identify the individual study population member to respond to the interviewer. Once identified, no substitution is allowed by the interviewers (Hess, 1985). If the identified individual cannot be reached or refuses to respond, the case is considered missing data and contributes to the NONRESPONSE BIAS for the study. With quota sampling, a refusal leads to an interviewer identifying another individual with the necessary characteristics and interviewing the newly identified person to fill his or her quota. Because quota samples do not begin with lists of the study population or maintain records of refusals, it is impossible to analyze the potential bias in these samples. However, many studies and analytical procedures have been conducted

to estimate the bias stemming from nonresponse on surveys using probability samples (Keeter, Miller, Kohut, Groves, & Presser, 2000).

SAMPLING METHODS RELATED TO QUOTA SAMPLING

Although quota samples are relatively rarely used in contemporary social science, there are several types of sampling methods that are similar and are used. One type of probability sampling, proportionate STRATIFIED SAMPLING (Henry, 1990; Kish, 1965), is sometimes confused with quota sampling because both can precisely replicate a few demographic characteristics of the study population. However, proportionate stratified sampling uses a random selection procedure, rather than interviewers, to select respondents. To use proportionate stratified sampling, a list of the study population that contains information about the variables that are used for MATCHING must be obtained. All of the individuals are then placed into mutually exclusive groups, or strata, based on the key variables. The sampling fraction for the study—that is, the percentage of the population that is to be drawn for the sample ($f = n/N$, where f is the SAMPLING FRACTION, n is the sample size, and N is the population size)—is applied to each stratum to determine the sample size for each stratum ($n_s = fN_s$). Finally, a random process is used to select the required number of respondents from each stratum (n_s). This process ensures that the sample selected and the study population will match on the key variables, and interviewer selection is removed as a form of bias.

In order to take advantage of the speed and low cost of quota samples, some survey researchers tried to combine quota sampling and proportionate stratified sampling into a technique that is known as “probability sampling with quotas” (Sudman, 1976). This

technique assigns interviewers to conduct interviews within specific neighborhoods, usually in a specific block, at specific times, and requires them to match a few demographic characteristics of the census tract that contains the specified block. With more controls and recordkeeping on the part of interviewers, biases due to refusals can be estimated and reduced.

Another use of quotas in sampling is to establish limits on the number of interviews from each stratum in a stratified sample. Often, this is done to oversample underrepresented groups, such as racial-ethnic minorities. All respondents are chosen by probability sampling methods, but once the quota for a stratum is reached, interviews with members of that stratum are curtailed after information identifying them as stratum members is obtained.

Quotas are also used in qualitative studies and other studies where an objective is to obtain data from a heterogeneous group of study population members. Although these studies yield biased population estimates, they increase the possibility that contrasting cases will be included in the study, and the differences between these cases can be explored through the research.

—Gary T. Henry

REFERENCES

- Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.
- Hess, I. (1985). *Sampling for social research surveys 1947–1980*. Ann Arbor, MI: Institute for Social Research.
- Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125–148.
- Kish, L. (1965). *Survey sampling*. New York: Wiley.
- Sudman, S. (1976). *Applied sampling*. New York: Academic Press.

R

R

The *R* is the symbol for a MULTIPLE CORRELATION COEFFICIENT. The bivariate correlation coefficient correlates two variables, say, *X* and *Y*. The multiple correlation coefficient is the correlation between two or more INDEPENDENT VARIABLES— say, X_1 , X_2 , and X_3 —and DEPENDENT VARIABLE *Y*. In this example, it would tell us how all three of these independent variables, taken together, correlate linearly with *Y*. The *R* ranges in theoretical value from .00 to 1.00. A high number indicates a high degree of linearity. A low number indicates a low degree of linearity, but does not rule out a high degree of NONLINEARITY. It should be noted that *R* is also used to label a statistical computer program that bears many similarities to S-plus and is gaining popularity in the social sciences.

—Michael S. Lewis-Beck

r

The *r* is the symbol for the PEARSON CORRELATION COEFFICIENT, as estimated in a SAMPLE. It is important to distinguish it from the population value, rho (ρ). It is the leading measure of correlation between two QUANTITATIVE VARIABLES. It assesses the strength of linear relationship, and goes from +1.0 or -1.0 to .00.

—Michael S. Lewis-Beck

RANDOM ASSIGNMENT

Random assignment refers to a way in which subjects are allocated into groups in an EXPERIMENT. One group is typically the CONTROL GROUP, and there can be one or more experimental groups. This way, the systematic effects of all other variables cancel out, and we can better assess the effect of the TREATMENT variable. Random assignment also makes it possible to make use of statistical inference methods and draw conclusions from the population from which the data came.

Many of the main ideas for how to run experiments originated with the British statistician Sir Ronald A. Fisher. He worked extensively with agricultural experiments. With many plots of land available for experiments, the question became how one would allocate various plots to the control and experimental groups and minimize the effect of variables other than the treatment variable. He wanted the assignment to be such that there would be no underlying factors that could explain the outcomes of the experiment. The effects of all variables other than the experimental variable should cancel out. The effects of these variables can never be completely eliminated, but their effects should not have a systematic presence in the data. Whatever effect there would be of all other variables gets absorbed in the residual variable.

Fisher's answer to this problem was to allocate the various elements (plots of land, people, animals, plants, etc.) to the different treatments in a random fashion. A random assignment means allocating elements to treatments such that there are no systematic effects of all variables other than the experimental

variable(s). When there is a control group and only one experimental group, one way to perform the random assignment is to toss a coin for each element. A “tail” would assign an element to the control group and a “head” to the experimental group. After the assignment is finished, there would be about the same number of elements in each of the two groups, and there would be no systematic effect of any other variables on the outcomes of the experiments.

A coin toss is a rather primitive way of assigning elements to treatments. Another way is to use a table of RANDOM NUMBERS. After a random starting point in the table, one can make use of whether a number is an odd or an even number. Many computer software packages generate random numbers that can be used for the random assignment of elements to control and experimental groups.

It is also possible to have a random assignment of elements when there is more than one experimental group. If Fisher planned to compare four different types of grain, plots of land would be assigned to the four experimental groups and a control group in a similar, random fashion.

Random assignments are used for the same purpose as random SAMPLING methods in observational studies. Random assignments and random sampling are needed in order to perform statistical HYPOTHESIS TESTING and the construction of CONFIDENCE INTERVALS.

—Gudmund R. Iversen

REFERENCES

- Brown, S. R., & Melamed, L. E. (1998). *Experimental design and analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-074). Thousand Oaks, CA: Sage.
- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Fisher, R. A. (1935). *The design of experiments* (7th ed.). New York: Hafner.

RANDOM DIGIT DIALING

Random digit dialing is a sampling technique used in TELEPHONE SURVEYS. Instead of using telephone directories that may be incomplete or out of date, random numbers are used to constitute telephone numbers. The digits relating to the target areas are

not randomly arrived at but are simply tacked onto the beginning of the randomly generated numbers. In the United States, this generally means adding a randomly generated four-digit code to the three-digit local exchange.

—Alan Bryman

RANDOM ERROR

Randomness lies at the heart of STATISTICAL INFERENCE. It is based on the notion that if an observation or a measurement is repeated, then most of the time, a different value is observed.

Randomness exists when it is not possible to predict the outcome of the next observation, either in a SURVEY or an EXPERIMENT. It follows that, in the presence of randomness for a variable, there will be RANDOM VARIATIONS in the observations, and the variable becomes a random variable. The magnitude of such randomness is measured by the random error.

The variation in the observations of a random variable leads to variations in any SAMPLE statistic computed from the data. Thus, a sample MEAN will vary from sample to sample, just as the percentage of voters who support a candidate varies from one sample of voters to another.

The variation in a sample statistic from one sample to the next can be seen when multiple samples are taken. A particular polling organization may not do repeated sampling, but prior to any major election, several polling organizations do their own studies, and the percentages they report vary from one study to the next. One of the reasons they vary is because of the randomness inherent in any sampling procedure. It is also possible to simulate such variation with the proper statistical software. Simulating samples of size $n = 1,000$ and a probability of 0.5 of drawing a 1 or a 0, these are the first few sample percentages of draws that are 1: 50.4, 50.4, 48.4, 51.4.

Such variation is known as the random error associated with the percentage. It is unfortunate that the word ERROR is associated with this variation, because there is nothing “wrong” with such variation. But to the extent that there exists one true percentage value in the population, any deviation of a sample percentage from this true POPULATION parameter can be thought of as an error.

HISTORICAL ACCOUNT

Physicists have long been concerned with how to measure characteristics of physical objects. Even repeated measures of the length of a yardstick result in different measurements, even though we realize that the yardstick has only one, fixed length (given fixed temperature, etc.). Physicists finally accepted this variation, and they attached the name “error” to the difference between an observed value and the true, underlying value. Statisticians carried on with the term when they observed different values of a statistic from different samples, even though this variation is due to sampling effects.

APPLICATIONS

The existence of the random error leads to the question of what the magnitude is of this error. If the magnitude of the random error is known, then it is possible to state whether a particular sample statistic exceeds the random error or not. How much would we expect a sample percentage to vary around the true parameter value? Statisticians have developed formulas for the computation of magnitudes of the random error. If such formulas do not exist, it is often possible to simulate many repeated samples with the proper software and thereby compute the magnitude of the random error.

EXAMPLES

One way to measure the magnitude of the random error is to find the STANDARD DEVIATION of a sample statistic across many different samples. Such a standard deviation is mostly known as the STANDARD ERROR of a sample statistic, to distinguish it from the standard deviation we compute from a set of single observations. For a sample of size 1,000 from a population split 50/50, the standard error of the sample percentage becomes 1.6. In most cases, two times the standard deviation, or two times the standard error, will include most of the data. Twice 1.6 equals 3.2, and most of the sample percentages from various samples should fall in the range from $50 - 3.2 = 46.8$ to $50 + 3.2 = 53.2$.

However, we are interested most often in whether a particular sample statistic lies within the random error inherent in the sampling procedure, or whether the sample statistic lies further away from the

presumed value of the population parameter. Suppose we hypothesize that a population percentage equals 50%. From a sample of 1,000 observations, the observed sample percentage equals 56.7%. Does 56.7% lie within the range around 50% produced by the random error, or is 56.7% beyond the random error produced by the sampling procedure?

We can now conclude that the sample percentage lies beyond the error that can be expected by the sampling error. We could change the sample percentage to a value of the standard normal variable. With a value larger than 1.96, we conclude that the deviation of the sample statistic from the hypothesized population parameter exceeds what can be expected on the basis of the random error itself. Here, $z = (56.7 - 50.0)/1.58 = 4.24$. For less than once in 10,000 different samples would we observe such a large value or a larger value of the standard normal variable. From this, we conclude that the hypothesized value of 50% for the population in its support of the candidate must be wrong. For the observed sample value to lie within the random error of the population percentage, the population percentage must lie closer to the observed sample value of 56.7%.

Any sample statistic will have its associated random error, as expressed in its standard error. We have some control of the magnitudes of the random errors from the way the sample is collected and which variables are used in a particular study. One way to make the random error smaller is to have a larger sample. The number of observations in the sample often figures directly into the computation of the standard error of a sample statistic. One drawback is that the sample size often enters the formulas through its square root, and it takes much larger sample sizes to get the desired effect. Because of the square root, we need a sample four times as large to get half the standard error. In MULTIPLE REGRESSIONS, COLLINEARITY will increase the standard error of the regression coefficients.

—Gudmund R. Iversen

REFERENCES

- Agresti, A., & Finley, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Iversen, G., & Gergen, M. (1997). *Statistics: The conceptual approach*. New York: Springer-Verlag.
- Moore, D. S., & McCabe, G. P. (1998). *Introduction to the practice of statistics*. New York: W. H. Freeman.

RANDOM FACTOR

Factor is another word for *variable*. Factor is commonly used in the running and analysis of data from EXPERIMENTS. For experiments, the word *treatment* is also used instead of factor or variable. For observational data, TREATMENT is often not an appropriate term. In a study of religious affiliation, treatment would not be a good term for a nominal variable with categories Catholic, Jewish, Protestant, other. It is not as if people are treated with different religions the way a plot of land is treated by a fertilizer.

However, for an experimental study of the effect of a NOMINAL VARIABLE on a QUANTITATIVE VARIABLE, the term *treatment* and the term *factor* are interchangeable, and both terms are used. Such a study would call for ANALYSIS OF VARIANCE, and because of its historical connection with experiments, *treatment* or *factor* are both used instead of the more neutral term *variable*. For such a study, the INDEPENDENT VARIABLE (treatment, factor) is thought of as having different *levels* instead of simply different *values*.

A factor in analysis of variance can be one of two kinds: *fixed* or *random*. A factor is considered fixed if we make use of all the possible categories of the factor. Gender, for example, would be a fixed factor when we use female and male as the two categories. For most purposes, there are no other categories for this variable. On the other hand, a factor is random if we have data for only some of the categories. In a study of differences among the 50 states, we can first choose a random sample of states, say, 10, and then collect data on the DEPENDENT VARIABLE from a sample from each of those chosen states.

Random factors are typically variables with such a large number of categories that it is not feasible or economical to collect data on the dependent variable for each of the categories. By first taking a random sample of categories and limiting the study to those categories, the study becomes more manageable.

An analysis of variance using a fixed factor as the independent variable is denoted as a MODEL I ANOVA. Similarly, an analysis using a random factor is denoted as a MODEL II ANOVA. In the case of only one factor, the computations of sums of squares and *F* VALUES are identical for the two types of models. But for two or more factors, it makes a difference in the computations whether we use a Model I or Model II. For example, in a two-way analysis of variance, in a Model I analysis, we

divide the row, column, and interaction mean squares by the residual mean square to get *P* VALUES. But in a Model II analysis, we divide the row and column mean squares by the interaction mean square to get *p* values.

It is also possible to have a mix of factors, some fixed and some random. Because it typically makes a difference whether a factor is classified as fixed or random, it is important to make certain one has computer software that takes this difference into account.

—Gudmund R. Iversen

REFERENCES

- Agresti, A., & Finley, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters*. New York: Wiley Interscience.
- Jackson, S., & Brashers, D. E. (1994). *Random Factors in ANOVA* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-098). Thousand Oaks, CA: Sage.

RANDOM NUMBER GENERATOR

A random number generator is any process that generates a sequence of values that is in some sense unpredictable. Random number generators are widely used in SIMULATION, Monte Carlo statistical techniques, and cryptography. True random numbers are generated by a natural process such as radioactive decay or atmospheric noise. However, most applications use pseudo-random numbers, which are generated by a deterministic equation that produces a sequence of values that has statistical properties similar to those of a true random sequence. Pseudo-random number generators are incorporated into most computer languages and in numerical software such as spreadsheets and statistical packages.

The most common technique for generating pseudo-random number generators is the linear congruent generator

$$y_{n+1} = (ay_n + b) \bmod m.$$

For example, in the ANSI C computer language, the `rand()` random number function has $m = 2^{31}$; $a = 1103515245$, and $b = 12345$. For reasons of computational efficiency, m is usually a power of 2; a and b are chosen experimentally on the basis of

the characteristics of the sequences produced by the equation. The linear congruent function generates integers in the range $[0, m - 1]$; if y_n is divided by $m - 1$, a sequence of uniformly distributed random numbers in the interval $[0, 1]$ can be generated. Uniform random numbers can, in turn, be used to generate random numbers for other PROBABILITY distributions, such as the normal distribution, by using specialized ALGORITHMS or by inverting the appropriate cumulative probability distribution function.

Because the linear congruent generator uses only y_n to determine y_{n+1} , once a value occurs more than once, all subsequent values in the sequence will be repeated in the same sequence that was generated after the first occurrence of that value. The number of values that can be generated without repetition is called the “cycle” or “period” of the generator. This is necessarily less than m , and in a poorly constructed generator, the period may be quite short.

An alternative approach is called the lagged Fibonacci generator, which has the form

$$y_{n+1} = (y_{n-i} + y_{n-j}) \bmod m$$

for some positive integers i and j . Fibonacci generators can have considerably longer periods, on the order of 10^{200} to 10^{500} .

The initial value used in a linear generator is called the “seed.” Different sequences will be generated depending on this value. In some applications, the same seed is used repeatedly so that the same sequence is generated each time; in other applications, the seed is set randomly, often using some randomly varying physical process, such as the number of seconds since the computer was turned on.

Developing high-quality pseudo-random number generators has proven to be a remarkably difficult problem in computer science; Knuth (1997) remains the definitive work on the problem. Generators can be assessed using a number of statistical tests; the key characteristics of a high-quality generator are a long period, absence of serial correlation between y_n and y_{n-k} , and uniformly filling of the interval $[0, m - 1]$. Finally, because applications often require the generation of billions of pseudo-random numbers, computational efficiency is important.

True random numbers are used in some specialized applications such as cryptography, where any pattern in the sequence might weaken the system. A number of Web sites exist that provide true random sequences based on physical processes such as

monitoring radioactive decay with a Geiger counter; see <http://www.random.org>. The decimal digits in irrational numbers such as pi or the square root of 2 also form true random sequences; the Web site http://antwpr.gsfc.nasa.gov/htmltest/rjn_dig.html provides sequences of several million digits from various irrational numbers. Unlike pseudo-random numbers, true random number sequences never cycle.

—Philip A. Schrodtt

REFERENCE

Knuth, D. E. (1997). *The art of computer programming, volume 2: Seminumerical algorithms*. Reading, MA: Addison-Wesley.

RANDOM NUMBER TABLE

We have all observed one or more numbers being drawn out of a hat in order to determine winners of prizes. For example, admission ticket stubs may be placed in a bowl and shaken before drawing a stub at random and reading off the prize-winning number. This is the function provided by a random number table. Table 1 is a brief excerpt of two lines from a page in a random number table. (The spaces between each set of five digits are present only for ease in reading and should be ignored in reading random numbers across or down the page.)

The entries in a random number table satisfy two basic properties:

1. Every digit from 0 to 9 has the same probability of occurrence (equally likely to occur).
2. The probability of any given number is not affected by the numbers that precede or follow it (independence).

One of the primary uses of a table of random numbers is for selecting a random sample for statistical analysis. Suppose we have a POPULATION of N items, and we wish to select a simple RANDOM SAMPLING of n items. We first assign a number label to each of the N items in the population, using the integers $0, 1, \dots, N - 1$. Those population items whose numbers are read from the random number table will become the elements in the sample. The first step is to choose a random starting point in the table. You could, for example, close your eyes and open the book

Table 1 Excerpt From a Random Number Table

77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952

to a random page and then point your finger to a starting point. From then on, we read successive digits across the page left to right (or right to left, or down the column, or diagonally, etc.). If $N - 1$ has x digits, we read x digits at a time in succession, ignoring any blank spaces.

For example, suppose we have a population of 683 items, and the above excerpt shows our random starting point, at the upper left-hand corner. Then, reading left to right, the numbers of the items to be included in the sample are 779, 210, 690, 711, 008, 427, 512, 775, 653, 498, and so on. Here, we have to throw out all numbers greater than 683. We keep reading successive three-digit numbers until we have obtained n items. The sample can be selected with replacement (if population item number 328 comes up more than once in the list of random numbers, that item is included in the sample more than once), or without replacement (if item 328 comes up more than once, it is thrown out after the first time).

It is important that every effort be made to ensure that the starting point in the table is different each time the table is used. One may wish to use the random number table itself to select the starting point. For example, we might read 56 from the excerpt above to indicate that we should begin on page 5 with the sixth digit.

The first random number table was produced in 1947. The RAND Corporation published the book titled *A Million Random Digits* in 1995; this book is still available for purchase and also appears online at <http://www.rand.org/publications/classics/randomdigits>. A paperback version was published in 2002.

—Jean D. Gibbons

RANDOM SAMPLING

Random sampling refers to SURVEY sampling designs that meet the criteria necessary to permit the use of randomization-based methods of STATISTICAL INFERENCE. The term *probability sampling* is synonymous. The defining characteristics of a random sampling design are the following:

1. The study population is fully defined.
2. The design is specified.
3. The design is objective.
4. The design is replicable.
5. Every unit in the population has a known probability of being selected.
6. All selection probabilities are nonzero.

These criteria warrant a little further explanation. The first seems straightforward but is often overlooked. It is necessary to define precisely the units of interest to the survey and the boundaries of the population. For example, it is not sufficient to say “all adults.” The definition should encompass geography, time, civil status, age range, and other relevant factors. A better definition might be “All people aged 18 or over whose sole or main residence as of 10 July 2003 was a private household in one of the 48 mainland states of the USA.” If the survey is carried out over a long period of time, we should also specify how we deal with the distinction between stock and flow.

Specification (Criterion 2) necessitates a clear and full description of the process that resulted in the selection of the sample.

Objectivity requires that each selection—and each stage in the selection process, because many designs are multistage—is governed by a random chance mechanism. It is not permissible for any subjective judgment to influence the selection of sample units. This is perhaps the heart of the definition of random sampling. A number of studies have shown that both experts and nonexperts produce samples that are seriously in error in terms of important characteristics if asked purposively to select a sample that they believe to be representative. Teachers of sampling also observe this phenomenon regularly in class exercises. Selection by means of a chance mechanism is the only way to avoid SYSTEMATIC ERROR (while also being able to measure variable error).

REPLICABILITY requires that another researcher, given identical circumstances, would be able to implement exactly the same design. This does not mean that he or she would necessarily replicate the *selected sample* (in fact, this would be extremely unlikely), but that he or she could replicate the *design*, that is,

reproduce the stages in the process and reapply the random mechanisms where appropriate.

Criterion 5 states that it is necessary to know the selection probability of every unit in the population. In fact, for purposes of estimation, it is necessary to know the probabilities of only those units that were actually selected—not of the ones that were not selected. The point is that it must, in principle, be possible to calculate the selection probability of any unit in the population. To meet this requirement, it is necessary first to have a design that is specified and objective, and, additionally, to have collected and recorded necessary information during the course of sample selection. For example, if the design involves selecting residential addresses from a list and subsequently selecting one person at random to interview at each address, it is necessary to record the number of eligible people at each sample address.

The final criterion is that every unit in the population must have a chance of being selected. It is not permissible for any units to be completely excluded from the selection process. If this happened, and the excluded units were in any way systematically different from the included units, a “coverage bias” would result.

Random sampling is widely considered to be the only acceptable approach for public sector and academic quantitative surveys from which descriptive statistics are to be produced. In the absence of other sources of systematic error (such as NONRESPONSE BIAS and measurement bias), it provides the basis for UNBIASED estimation. However, model-based approaches to INFERENCE argue that randomization is not necessary provided that sources of systematic error in the sample structure are recognized in the model (e.g., as covariates). Increasingly, the two approaches are combined by the use of model-assisted inference based on random sampling (Brewer, 2002).

Many government and academic social surveys have a requirement for random sampling. This is set out in protocols such as those for the European Labour Force Surveys (Eurostat, 1998), European Social Survey (ESS-CCT, 2001), and International Social Survey Programme (Park & Jowell, 1997). However, it is not uncommon to find surveys that purport to use random sampling but, in fact, diverge from the criteria in important ways—perhaps by allowing PURPOSIVE selection of sample areas at the first stage, or by using QUOTA SAMPLING within households or other randomly selected units.

A wide range of sampling designs is possible within the definition of random sampling, including the use of MULTISTAGE SAMPLING, proportionate or disproportionate STRATIFIED SAMPLING, CLUSTER SAMPLING, and SYSTEMATIC SAMPLING.

With random sampling, it is necessary to make considerable efforts to maximize survey response rates. Without this, SAMPLING BIAS may simply get replaced by nonresponse bias. This is sometimes seen as a disadvantage of random sampling, because the cost per survey response is typically much higher than with nonrandom methods (particularly in the case of face-to-face interviewing). The alternative to random sampling methods is purposive sampling methods, of which quota sampling methods are the most well known.

—Peter Lynn

REFERENCES

- Barnett, V. (1974). *Elements of sampling theory*. Sevenoaks, UK: Hodder & Stoughton.
- Brewer, K. (2002). *Combined survey sampling inference: Weighing Basu's elephants*. London: Arnold.
- European Social Survey Central Co-ordinating Team. (2001). *European Social Survey (ESS): Specification for participating countries*. London: Author.
- Eurostat. (1998). *Labour force survey: Methods and definitions* (1998 ed.). Luxembourg: Author.
- Moser, C. A., & Kalton, G. (1971). *Survey methods in social investigation* (2nd ed.). Aldershot, UK: Gower.
- Park, A., & Jowell, R. (1997). *Consistencies and differences in a cross-national survey*. London: SCPR.

RANDOM VARIABLE

In statistics, a variable can be *fixed* or *random*. A random variable is a variable where we cannot predict the value for an observation or necessarily get the same value when we repeat the observation. A fixed variable does not change value across repeated measurements. The distinction between the two types can be rather fluid, but it is usually clear from the context whether a variable is fixed or random. Statistical inference (HYPOTHESIS TESTING and CONFIDENCE INTERVAL estimation) applies only to random variables.

We would think that the measure of the length of a yardstick is a fixed variable, because the length does not change. But physicists established early on that when such a stick is measured many times, the results were

not the same from measurement to measurement, even though the stick stayed the same. Thus, the length is a random variable, and the difference between the actual length, whatever it is, and the observed length is due to MEASUREMENT error.

In the study of relationships between variables, the INDEPENDENT VARIABLE(S) is often considered a fixed variable, whereas the DEPENDENT VARIABLE is considered a random variable. In a study of the relationship between education (measured as number of completed years in school) and income, education would be considered a fixed variable and income would be considered a random variable. The education of a person would be the same no matter how often it is measured; people with 16 years of education would continue to have that value (unless they go on to graduate school). People with the same education would have different incomes, and there would be a spread of income values around a mean income for each particular educational group.

In a REGRESSION analysis, usually there are variations in the dependent, random variable around the regression line for each value of the independent, fixed variable. This is the reason we minimize the vertical SUM OF SQUARES of the dependent variable around the regression line, instead of minimizing the sum of squares in the horizontal or perpendicular direction in the scatterplot. Minimizing the sum of squares of the dependent variable enables us to bring the machinery of regression analysis to these DATA to find results such as the regression line, the CORRELATION coefficient, and *P* VALUES. All of the formulas for ordinary regression analysis in popular statistical packages are based on the independent variable(s) being a fixed variable(s) and the dependent variable being a random variable.

What happens when both the independent and the dependent variables in a regression analysis are considered random variables? There are times when it makes sense to consider both independent and dependent variables as random variables. In this situation, the mathematical derivations of the necessary formulas become much more difficult and do not lead to the same simple formulas as with a fixed, independent variable and random, dependent variable. Mostly, however, people overlook the problem and use the formulas for ordinary regression anyway, presumably without too much damage to the analysis and the result.

—Gudmund R. Iversen

REFERENCES

- Iversen, G., & Gergen, M. (1997). *Statistics: The conceptual approach*. New York: Springer-Verlag.
- Moore, D. S. (1997). *Statistics: Concepts and controversies*. New York: W. H. Freeman.
- Utts, J. M. (1996). *Seeing through statistics*. Belmont, CA: Duxbury.

RANDOM VARIATION

The study of variation is a central idea in statistics. Variation means that when we measure something over again, we get a different result, and we cannot predict the outcome of any future observation. Any time we have a RANDOM VARIABLE, there is, by definition, variation attached to the variable.

Two questions arise: (a) How do we measure variation? and (b) How much of the variation in a variable is produced by other variables? A simple way to measure the amount of variation is to compute the standard deviation or some related STATISTIC. The contribution to the variation by other variables is usually measured by sums of squares or MULTIPLE CORRELATIONS.

One way to understand the magnitude of variation in a variable is to take the variable as a DEPENDENT VARIABLE and introduce one or more INDEPENDENT VARIABLES, say, in a REGRESSION analysis. Such an analysis results in various sums of squares. The total sum of squares, the numerator in the VARIANCE of the dependent variable, gives a measure of the total variation in the dependent variable. There are two reasons why the values of the dependent variable are different. One is that units have different values of the independent variables; the other reason is that there are effects of variables other than the independent variables. The combined effects of all these other variables are known as the effect of the RESIDUAL variable (or ERROR variable).

Some of the total variation can be traced to the independent variables. We can use the estimated regression equation to compute a set of predicted values of the dependent variable. Why are these predicted values different from each other? They are different because they were computed from the independent variables, and observations have different values on the independent variables. Thus, the variation in the predicted values of the dependent variable is due to the independent variables. The amount of this variation is measured

by computing the REGRESSION SUM OF SQUARES. It gives the effects of the independent variables on the dependent variable in terms of how much variation in the dependent variable is produced by the independent variables.

The difference Total Sum of Squares minus Regression Sum of Squares gives the magnitude of the effects of the residual variable, or the amount of the total variation produced by all other variables. This difference is known as the RESIDUAL SUM OF SQUARES or the error sum of squares (even though there is nothing wrong with it). This sum of squares can also be found from taking the difference between observed and predicted values of the dependent variable and adding the squares of these differences.

Each time we add another variable with its parameter to the model, the residual sum of squares stays the same or gets smaller. But with more parameters, the model becomes less parsimonious. In the extreme, by including as many parameters as there are observations, the residual sum of squares equals zero. However, nothing has been gained, because we have only replaced n quantities (data points) by n other quantities (parameters).

—Gudmund R. Iversen

REFERENCES

- Iversen, G., & Gergen, M. (1997). *Statistics: The conceptual approach*. New York: Springer-Verlag.
- Moore, D. S. (1997). *Statistics: Concepts and controversies*. New York: W. H. Freeman.
- Utts, J. M. (1996). *Seeing through statistics*. Belmont, CA: Duxbury.

RANDOM WALK

A random walk is a STOCHASTIC process sometimes used as a model for TIME-SERIES data. Perhaps the best known application of a random walk is to stock market prices on successive days.

Suppose that $[u_t]$ is a sequence of RANDOM VARIABLES that is mutually independent and identically distributed. As such, $[u_t]$ has a constant mean μ and variance σ^2 . The covariance between u_t and u_{t+k} , $k = \pm 1, \pm 2, \dots$, is equal to 0. A process that would produce random variables with these properties is sometimes called a “purely random process” or WHITE NOISE.

Drawing heavily on Chatfield’s (1996) fine exposition, a set of random variables $[y_t]$ is viewed as a random walk if

$$y_t = y_{t-1} + u_t. \quad (1)$$

In practice, the time series is assumed to start at 0 when $t = 0$. Then, $y_1 = u_1$ and $y_t = \sum_{i=1}^t u_i$.

It follows that $E(y_t) = t\mu$ and $\text{Var}(y_t) = t\sigma^2$. Because the mean and the variance change with t , a random walk is not stationary. However, taking the first difference, one gets $(y_t - y_{t-1}) = u_t$, the purely random process with which we started.

A random walk is a special case of a first-order AUTOREGRESSIVE process,

$$y_t = \alpha y_{t-1} + u_t, \quad (2)$$

where $\alpha = 1$. The links between a random walk and Autoregressive Integrated Moving Average (ARIMA) models more generally are discussed in Box, Jenkins, and Reinsel (1994).

—Richard A. Berk

REFERENCES

- Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). New York: Prentice Hall.
- Chatfield, C. (1996). *The analysis of time series: An introduction*. New York: Chapman & Hall.

RANDOM-COEFFICIENT MODEL.

See MIXED-EFFECTS MODEL;
MULTILEVEL ANALYSIS

RANDOM-EFFECTS MODEL

ANALYSIS OF VARIANCE (ANOVA) models can be used to study how a dependent variable Y depends on one or several factors. A factor here is defined as a categorical independent variable; in other words, an explanatory variable with a nominal LEVEL OF MEASUREMENT. In a random-effects ANOVA model (also called a Type II ANOVA model), the effects of the categories of each factor are modeled as random variables; one could say that this is a model with random parameters. More specifically, when a given factor has

values $1, 2, \dots, K$ in the data set, and for case i , the value of this factor is indicated by $k(i)$, then a random effect of this factor for case i is defined by the term $U_{k(i)}$ in the linear model for Y_i , where it is assumed that $U_1, \dots, U_K$ are independent and identically distributed RANDOM VARIABLES. Usually, a normal distribution is assumed for these random variables. Such a model is appropriate if the categories of this factor are regarded as a random sample from some population and if the statistical inference aims at drawing conclusions that hold for this population. This contrasts with FIXED EFFECTS MODELS, which are described in another entry. Random-effects ANOVA models also are called VARIANCE COMPONENTS MODELS.

Two examples of studies where random-effects models can be appropriate are (a) educational studies about educational achievement of pupils where teachers are an important factor, and the researcher is primarily interested in drawing conclusions that apply to the population of teachers rather than specifically to the teachers included in the present study; and (b) studies in the framework of GENERALIZABILITY THEORY and INTERRATER RELIABILITY, where ratings on individuals are obtained from each of several observers, and one of the interests is in the contribution of the observers (drawn randomly from some population of observers) to the variability of the measurement obtained by averaging the ratings of one individual over the observers.

Random effects occur similarly in extensions to other, more complicated models, such as GENERALIZED LINEAR MODELS or nonlinear regression models. Their definition, then, is the same: The effects of the categories of each factor are modeled as random variables.

The term *random-effect models* is reserved for models in which each factor has random effects. Models containing some random as well as some fixed effects are called MIXED-EFFECT MODELS, and these are used more frequently than pure random effect models. The random effects as defined above can be regarded as random main effects of the levels of a given factor. A generalization is a random coefficient of a numerical explanatory variable (see RANDOM-COEFFICIENT MODELS). For hierarchically nested factors, models including random coefficients as well as fixed effects are also called HIERARCHICAL LINEAR MODELS. These models are the basis of most procedures of MULTILEVEL ANALYSIS.

—Tom A. B. Snijders

REFERENCES

- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Jackson, S., & Brashers, D. E. (1994). *Random factors in ANOVA* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–098). Thousand Oaks, CA: Sage.
- Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. W. (1996). *Applied linear regression models*. New York: Irwin.
- Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.

RANDOMIZED-BLOCKS DESIGN

As the outcome of an EXPERIMENT, the values of the DEPENDENT (response) VARIABLE will differ from each other from one experimental unit to the next. It is this variation in the values that gives us clues about the effects of the INDEPENDENT (treatment) VARIABLES and perhaps other variables. The variation in the values of the response variable is typically measured by subtracting the mean from each observation, squaring all the differences, and adding all the squares. This procedure gives the total sum of squares, $TSS = \sum (y - \bar{y})^2$.

Experiments can be set up in different ways, using different designs. The goal of any design is to isolate the effects of the TREATMENT variables from other variables. We want the variation in the values of the response variable to be due to the treatment variables and not to other, extraneous variables. In a randomized-blocks design, similar units are gathered in groups called blocks. Both experimental and CONTROL GROUPS contain blocks.

EXPLANATION AND EXAMPLE

For example, a psychologist may have a suspicion that the gender variable affects the outcome of an experiment in perception. Without the use of blocks, the experiment consists of an experimental and a control group. After the experiment, there is a mean score for the experimental group and a mean score for the control group. The impact of the treatment can then be seen in how different these two means are from each other and from the overall mean. One way to measure this impact is to compute the treatment sum of squares.

The difference between the total sum of squares and the treatment sum of squares becomes the residual sum of squares. This sum gives the combined effects of all variables other than the treatment variable, including

gender. To establish STATISTICAL SIGNIFICANCE, it is common to divide the treatment and the residual sums of squares by their appropriate degrees of freedom to get the mean squares. They are used to find the value of the statistical F variable, where F = treatment mean square/residual mean square. The larger the F , the smaller the P VALUE and the more significant the results are.

One way to get a large value of F is to have a large numerator, and another way is to have a small denominator. There is not much that can be done with the numerator, but it may be possible to reduce the denominator. Part of the reason the values of the dependent variable are different may be because people are of different genders. If it were possible, somehow, to take out the magnitude of this source of variation in the residual sum of squares, then we would get a smaller, new residual sum of squares and thereby a smaller denominator for F , which again would give a larger value of F .

Here is where the idea of blocking comes in. By identifying the gender of each subject, it is possible to arrange the mean perception scores for each gender and for treatment/control groups in a table (see Table 1).

Table 1 Mean Scores for Blocks and Treatments

Block	Treatment Group	Control Group	Mean
Female	$\bar{y}_{ft}$	$\bar{y}_{fc}$	$\bar{y}_f$
Male	$\bar{y}_{mt}$	$\bar{y}_{mc}$	$\bar{y}_m$
Mean	$\bar{y}_t$	$\bar{y}_c$	$\bar{y}$

Such an arrangement of the means does not change the computation of the treatment sum of squares. It still simply involves the means for the treatment and control groups. Thus, the numerator for the F does not change from how it was computed without blocking.

But it is now possible to compute a block sum of squares. It is found from the means of the two gender groups, the number of observations in each block, and the overall mean. This sum of squares very much resembles the treatment sum of squares above. It is found with the following formula: Block Sum of Squares = $\sum n_f = (\bar{y}_f - \bar{y})^2 + \sum n_m (\bar{y}_m - \bar{y})^2$. Next, we subtract this sum of squares from the old residual sum of squares to get a new residual sum of squares, smaller than the old one. By properly adjusting for the various DEGREES OF FREEDOM, we get a new and smaller denominator for F than what we had without blocking. That results in a larger F and a smaller p value.

The key here is that we have been able to “purify” the residual sum of squares. The residual variable can be seen as the net effect of all other variables, the blocking variable is one of them, and we have been able to take out the effect of this one variable. Therefore, the remaining variation in the residuals is, most of the time, smaller than it was. The residuals are now found as the difference between each observation and the appropriate doubled subscripted mean above; this is instead of the difference between an observation and the mean of either the treatment or the control group, when blocks were not taken into account.

HISTORICAL ACCOUNT

Much of the thinking behind the planning of experiments goes back to the pioneering ideas by Sir Ronald A. Fisher on his work on agricultural data. He noted that on a field of land divided up into experimental units, units in one part of the field may be different from units in other parts of the field because of things such as difference in altitude and difference in amount of sunlight. If these factors were not in some way taken into account, they could well influence the amount of yield from the different units and affect the variation in the values of the yield. One way around such a problem is through the use of a randomized-blocks design. Instead of simply dividing the units into experimental and control groups, first it would make sense to group similar land units into what he called blocks.

APPLICATIONS

Without blocking, the analysis consisted of a simple, one-way analysis of variance. With blocking and the mean arranged as in the table above, the analysis consists of a two-way analysis of variance but without taking into account any idea of a treatment-block interaction effect.

—Gudmund R. Iversen

REFERENCES

- Agresti, A., & Finley, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Moore, D. S., & McCabe, G. P. (1998). *Introduction to the practice of statistics*. New York: W. H. Freeman.

RANDOMIZED CONTROL TRIAL

A randomized control trial (RCT) is a type of evaluation research design. RCTs address questions about whether, and to what extent, interventions might “work.” A trial is a test, the term *controlled* indicates the inclusion in the study design of one or more comparison groups, and RANDOMIZED refers to a method of using random allocation to generate intervention and comparison groups. The use of random allocation is the only characteristic that distinguishes RCTs from other prospective evaluation designs. The RCT’s role is to reduce the chances of being misled about intervention impact because of unmeasured factors associated with other ways of selecting intervention and comparison groups. Randomization also permits legitimate statistical statements about the probability that study results are a matter of chance.

WHY RCTs ARE USEFUL

Assessing the effectiveness of different types of interventions is a major challenge for researchers, scientists, policymakers, and the public generally. Does intensive social work input for children with behavior problems actually help these children and their families? What is the impact on children’s literacy of computer-assisted learning? Does introducing criminals to their victims secure lower reoffense rates? Do screening programs for breast or prostate cancer work in reducing deaths from these diseases? Answering such questions requires the use of a research design that permits the impact of the intervention in question to be disentangled, so far as is possible, from other influences, including the social characteristics of the intervention recipients and whatever else is going on in the social context at the time. RCTs can provide rigorous evidence to inform policy and practice decisions about the relative effectiveness of different interventions. As in any research, a good design is only optimally useful if it addresses an important question and is well operationalized.

THE HISTORY OF RCTs

Efforts to control biases in assessing the effects of interventions go back at least three centuries, include both medical and social science, and are based on the important recognition that even well-intentioned

interventions may do harm. Physicians from the 17th century on recommended casting lots to determine which patients would receive a particular treatment. This recognition of the need to compare like with like gathered momentum during the 18th century, such as naval surgeon James Lind’s famous study of scurvy, in which patients with the same condition were assigned to different treatments, and those who got oranges and limes recovered most quickly. The use of alternate or “random” allocation to generate comparison groups became more common after about 1920, with the first detailed descriptions of random allocation processes occurring in the 1930s. The term *control* was used in experimental psychology from the 1870s on to mean a standard of comparison used to check inferences made from an experiment; it comes from “counter-roll,” a duplicate register made to verify an official account.

Although today, RCTs are mainly associated with health care, social scientists have been equally concerned with designing intervention evaluation so as to be able to answer impact questions reliably. In the 1920s, 1930s, and 1940s, researchers, particularly in North America, in education, social welfare, and public policy made extensive use of experimental research designs, and “experimental sociology” was an important field of study. This was followed by “the golden age of evaluation,” from the 1960s to the 1980s, when American evaluators, often with government backing, used RCTs to evaluate a wide range of public policy initiatives, including income maintenance and housing allowance schemes, programs of support for disadvantaged workers, and interventions for former prison inmates.

RCTs AND OTHER RESEARCH DESIGNS TO ASSESS EFFECTIVENESS

An RCT tests the effects of an intervention by comparing defined outcomes in one population that is offered an intervention with those in another population that is not offered it. Potential recipients of an intervention are assigned to intervention or control groups using some method of random allocation (chance), such as tossing a coin or using a computer randomization program. Either individuals or groups (such as schools, hospitals, or communities) can be randomized (“cluster” randomized trials). Cluster RCTs may be particularly useful in the social sciences.

Empirically, results of RCTs usually differ from those of nonrandomized or uncontrolled studies (for

example, the common design of pre- and posttest on one group only). RCTs often yield lower estimates of effectiveness than other types of studies. There are many examples in social science and policy where the results of RCTs have been different from those of observational studies. For example, an RCT of day care for young children, the Perry Preschool Project, has shown with a follow-up of 27 years that this intervention can have long-term beneficial effects on such measures as educational achievement, unintended pregnancy, and criminal behavior, and this is in contrast to the findings of observational studies, many of which suggest deleterious effects of day care.

CURRENT DEBATES ABOUT RCTs

Most recent work on the methodology of RCTs has been carried out in health care by researchers allied with the Cochrane Collaboration (named after Archie Cochrane, an epidemiologist who argued strongly for the synthesis of research evidence based on RCTs). A sibling network, the Campbell Collaboration, was set up in 2000 to apply the same principles to the analysis of evidence about social interventions. This is named after Donald Campbell, a methodologist, whose 1966 book with Julian Stanley, *Experimental and Quasi-Experimental Designs for Research*, established the principles and procedures for the use of RCTs in social science.

Unfortunately, the association of the term “RCT” with medicine leads many social scientists to dismiss this method. Objections to the ethics of RCTs ignore the unethical nature of research designs that cannot properly answer the research question. They also fail to acknowledge the normal practice of uncontrolled experimentation, which ensures that people may be subjected to professional and other interventions with unknown, and possibly harmful, effects. The many thousands of RCTs conducted of social interventions show that this method is practical, even in complex social settings. A well-conducted RCT is a powerful method for assessing strategies for solving some of today’s most pressing social problems.

—Ann Oakley

REFERENCES

- Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.
- Campbell, D. T. (1988). *Methodology and epistemology for social science: Selected papers* (E. S. Overman, Ed.). Chicago: University of Chicago Press.
- Oakley, A. (2000). *Experiments in knowing: Gender and method in the social sciences*. Cambridge, UK: Polity.
- Orr, L. L. (1999). *Social experiments: Evaluating public programs with experimental methods*. Thousand Oaks, CA: Sage.

RANDOMIZED RESPONSE

When asked sensitive or threatening questions within a SURVEY context, respondents frequently lie, despite assurances of confidentiality. Indeed, advancing the cause of social science research is not, for most people, a sufficiently compelling reason to disclose embarrassing or even incriminating information.

Randomized response is a data collection strategy that incorporates lying of a sort into the survey instrument in order to protect respondents. That is, respondents are encouraged to give ambiguous answers based on some random process. Like the popular game show *Jeopardy*, the interviewer knows the respondent’s answer, but not exactly the question to which it is associated.

In its simplest form (Warner, 1965), a respondent is instructed to answer a sensitive question or its converse depending on the outcome of some randomizing device, such as the roll of a die. For example,

- | | |
|-------------------|--|
| Roll 1 or 2: | Have you used drugs in the past month? |
| Roll 3 through 6: | Have you abstained from using drugs in the past month? |

The catch is that only the respondent knows the roll of the die, so both a “yes” response and a “no” response may be either an admission or a denial of drug use, only the respondent knows for sure. Yet, in the aggregate, the prevalence of drug use can be estimated from the SAMPLE percentage of affirmative responses (some of which indicate drug use and the remainder claim abstinence).

The proportion of affirmative responses (π) is

$$\lambda = p\pi + (1 - p)(1 - \pi),$$

where p is the BINARY probability for the randomizing device (e.g., $p = 1/3$ for a 1 or 2 on a die) and λ is the desired prevalence parameter (e.g., the

percentage using drugs in the past month). We can then derive an estimate of the prevalence of the sensitive attribute under investigation from the sample proportion of affirmative responses,

$$\hat{\pi} = \frac{\hat{\lambda} + p - 1}{2p - 1}.$$

Although elegant in its simplicity, pairing a sensitive question with its converse may provoke respondent suspicion about statistical trickery. A more useful strategy, known as the unrelated question approach (see Horvitz, Shah, & Simmons, 1967), enhances respondent comfort by including a nonthreatening alternative. A respondent is instructed to answer a sensitive question or the alternative item depending on the outcome of the randomized device (say, a coin). For example,

Heads: Have you used drugs in the past month?
Tails: Is the final digit of your social security number odd?

In the aggregate, knowing that half the respondents would have been directed to answer the social security number question, and half of those would have responded "yes," we can filter the response distribution accordingly. In estimating the prevalence of drug use, we purge the observed response distribution of the expected number of affirmative responses to the nonsensitive question.

In the years since Warner's seminal paper, numerous alternative approaches using multiple questions, multiple groups, and various randomizing devices have been proposed (see Fox & Tracy, 1986, and Chadhuri & Mukerjee, 1987, for extensive discussions). The objective, of course, is to design a survey question that inoculates respondents in sensitive areas, that is credible, yet that is statistically efficient. The ambiguity created by the randomizing device and the use of multiple questions, designed to reduce evasive response bias, comes at a statistical cost in terms of precision.

The randomized response approach has also been extended to quantitative responses. For example, a respondent may be asked to respond to one of two questions based on drawing a colored ball from an urn:

Blue ball: How many days in the past week have you used drugs?
Red ball: What is the final digit of your social security number?

The mean of the observed responses (Z) is a weighted average of the mean frequency of the sensitive question (X) and the mean of the alternative question (Y), weighted by the probability (p) associated with the randomizing device. Thus, the estimated mean for the sensitive question (X) can easily be derived from

$$\hat{\mu}_X = \frac{\bar{Z} - (1 - p)\mu_Y}{p}.$$

The mean as well as percentage distribution of the sensitive question can be easily derived from the complete collection of responses, based on knowledge of the distribution of the nonsensitive characteristic and the randomizing device.

Across all designs, the randomized response strategy permits statistical estimates of aggregate parameters without revealing anything definitive about individual respondents. Even without individual-level scores, however, virtually any type of MULTIVARIATE technique usually calculated from individual scores can still be performed. CORRELATIONS and other statistics based on ASSOCIATION can be estimated by treating randomized response data as a special case of MEASUREMENT ERROR with known (or estimable) distributional properties.

In the years since its introduction, randomized response has been successfully applied in many areas of social science to obtain estimates of sensitive behaviors, such as academic cheating, abortion, organizational wrongdoing and whistle-blowing, the use of questionable business practices, sexual aggression, child molestation, AIDS-risky sexual behaviors, substance use, and other forms of criminal activity (see Daniel, 1993, for a review).

Despite its statistical appeal, the randomized response technique has not exactly been embraced by the applied research community. Although the technique is mathematically sound, concerns have been raised over respondents' understanding and trust in it. Yet several field validations have shown promising results. A number of studies have tested performance of the randomized response technique against known data as well as compared estimates obtained using randomized response with results of other relevant methods, such as direct questioning (see Armacost, Hosseini, Morris, & Rehbein, 1991; Stem & Lamb, 1981; Tracy & Fox, 1981; van der Heijden, van Gils, Bouts, & Hox, 2000). The randomized response technique has consistently been shown to outperform

conventional interviewing techniques, yielding higher estimates of occurrence of socially undesirable or illicit behavior.

Although the randomized response technique has been shown to be effective in reducing response bias (tendency to give untruthful answers), there has been mixed evidence as to whether this technique helps to reduce NONRESPONSE BIAS (refusal to participate). Respondents may feel more comfortable answering questions honestly when no one knows which question they are answering, but their perception of the technique as too complicated or laborious may be reason to refuse participation. In the final analysis, randomized response techniques may be most appropriate for relatively sophisticated populations—with respondents able to comprehend and appreciate the protection afforded them through the technique. Even then, some respondents may be as concerned about the impact of their responses, even in the aggregate, on the reputation of their organization or profession as they are about their own privacy.

There are some important considerations for successfully employing the randomized response approach whenever it is used. First, sample sizes must be relatively large to counterbalance the increased uncertainty (i.e., SAMPLING ERROR) that derives from the random error infused by the technique. Second, because of the increased error, the technique is not recommended for investigating rare events, particularly those with prevalence less than 10%. Despite these precautions, there are still certain social science inquiries so delicate that a method like randomized response is simply indispensable.

—James Alan Fox and Maria Tcherni

REFERENCES

- Armast, R. L., Hosseini, J. C., Morris, S. A., & Rehbein, K. A. (1991). An empirical comparison of direct questioning, scenario, and randomized response methods for obtaining sensitive business information. *Decision Sciences*, 22, 1073–1090.
- Chadhuri, A., & Mukerjee, R. (1987). *Randomized response: Theory and techniques*. New York: Marcel Dekker.
- Daniel, W. W. (1993). *Collecting sensitive data by randomized response: An annotated bibliography* (2nd ed.). Atlanta: Georgia State University Business Press.
- Fox, J. A., & Tracy, P. E. (1986). *Randomized response: A method for sensitive surveys*. Beverly Hills, CA: Sage.
- Horvitz, D. G., Shah, B. U., & Simmons, W. R. (1967). The unrelated question randomized response model. In E. D. Goldfield (Ed.), *Proceedings of the Social Statistics Section* (pp. 65–72). Washington, DC: American Statistical Association.
- Stem, D. E., Jr., & Lamb, C. W., Jr. (1981). The marble-drop technique: A procedure for gathering sensitive information. *Decision Sciences*, 12, 702–707.
- Tracy, P. E., & Fox, J. A. (1981). The validity of sensitive measurements: A comparison of two measurement strategies. *American Sociological Review*, 46, 187–200.
- Van der Heijden, P. G. M., van Gils, G., Bouts, J., & Hox, J. J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods & Research*, 28, 505–537.
- Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63–69.

RANDOMNESS

In statistics, randomness is the fundamental concept on which STATISTICAL INFERENCE is founded. With the proper presence of randomness, we can perform HYPOTHESIS TESTING and construct CONFIDENCE INTERVALS for unknown parameters.

Randomness exists when it is not possible to predict the outcome of an experiment or observation before it is performed.

The concept itself has a broader presence than how it is used in statistics. Simple examples of the presence of randomness include gambling, tossing a coin or a die, and playing cards. Before a coin is tossed, we cannot tell whether heads or tails will show. This leads to “coin toss” becoming a RANDOM VARIABLE with two values, unless we are exceptionally good at tossing coins. Most dependent variables in statistical analyses are random variables.

The opposite of a random variable is a variable where the values are fixed. Most INDEPENDENT VARIABLES are thought of as variables with fixed values. We consider education (say, years of schooling) as fixed. At the same time, income is considered a random variable. It is only with random sampling that we can use people with fixed years of schooling and study the DISTRIBUTION of the incomes for each school year group.

EXPLANATION

In STATISTICS, randomness is required for two things, the collection of DATA and the use of statistical inference. For OBSERVATIONAL data, the elements in a SAMPLE (people, counties, etc.) must be chosen

randomly from the larger POPULATION. A variety of sampling schemes can be used, such as SIMPLE RANDOM SAMPLING, STRATIFIED SAMPLING, and CLUSTER SAMPLING. It is only with random sampling that we can use statistical formulas for the computation of P VALUES or confidence intervals. Those formulas are derived mathematically under the assumption that the data are collected according to proper statistical sampling procedures. Indeed, most formulas in the popular statistical computing packages are only appropriate under the assumption of simple random sampling.

Randomness is also needed for EXPERIMENTAL data. Elements must be assigned to experimental and control groups in a random fashion. A variety of experimental designs exist that guarantee the proper randomness for the data.

HISTORICAL ACCOUNT

References to randomness are found as early as the Old Testament and Viking sagas. They refer to decisions being made by some random mechanism. More extensive studies of randomness started in the 17th century, when gentlemen gamblers wanted to establish rules for payoffs from games of chance with cards and dice. This led to mathematicians developing what we think of today as PROBABILITY theory.

Randomness has its foundations in philosophy of science, and it raises several issues. For example, it can be argued that there are degrees of randomness. The more knowledge we have about a phenomenon, the less random it becomes. In Bayesian statistics, this leads to posterior distributions with decreasing spread.

Also, a person really good at coin tossing may be able to toss a coin in such a way that we can accurately predict for each toss whether it will come up heads or tails. Similarly, in MULTIVARIATE statistical analyses, we can add variables in a REGRESSION analysis and thereby reduce the effect of the residual variable, which is thought to represent the randomness in the dependent variable. If we can add variables such that the MULTIPLE CORRELATION coefficient R becomes equal to 1, then there is no randomness left.

APPLICATION

In the rarified world of vacuum, there is no randomness associated with the time it takes an object to fall a certain distance, and physicists have an equation that relates these variables. With the absence of randomness, the variables are related as expressed in

a natural law. Such laws have often been the envy of social scientists, who have tried to relate variables in similar ways, so far without any success.

Randomness provides the foundation for probability theory. When we do not know the outcome of an experiment or an observation, we try to associate the various outcomes by probabilities. Indeed, it can be said that probability theory consists of the study of randomness. In turn, probability theory provides the foundation for statistical inference.

How do we create randomness and thereby collect random data? This is easier said than done. Samples have to be drawn using methods such as simple random sampling or maybe multistage stratified sampling, based on random numbers. An extreme example of the lack of randomness occurred in the first draft lottery during the Vietnam War, where the selection of birthdays was not random, and young men born later in the year became more likely to be drafted than those born earlier in the year.

EXAMPLES

How do we know that random numbers truly are random? What do we mean by a set of numbers being a random set of numbers? If random numbers were used only for the drawing of random samples, this may not be a very important question. But random numbers are also used in other connections, most importantly in cryptographic work for the construction of secret codes. Also, state lotteries make use of randomness, and anything less than random winning numbers would not be acceptable by the public.

This means that computer scientists and others have given much thought to how to create collections of random numbers as well as trying to determine what it means to have a truly random collection of numbers. The procedure should be such that if we generate a second series of random numbers, we should get a different series of numbers. But most computer programs depend on the so-called seed number to generate random numbers. If we use the same seed number a second time, we get the same collection of numbers, meaning they are not random. Also, each digit should appear about as often as the other digits. But if all sequences of digits are equally likely, we could get a sequence of, say, all 2s. Even though it was generated randomly, we would hesitate using such a sequence for anything.

Humans are not very good at predicting the future, even just the next few minutes. This implies that we find

ourselves living in a random world. To make sense of this randomness, we try to discover regularities in the randomness and thereby create some degree of order.

—Gudmund R. Iversen

REFERENCES

- Beltrami, E. J. (1999). *What is random? Chance and order in mathematics and life*. New York: Copernicus.
- Bennet, D. J. (1998). *Randomness*. Cambridge, MA: Harvard University Press.
- Keren, G., & Lewis, C. (Eds.). (1993). *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Lawrence Erlbaum.

RANGE

Range is the distance between the extreme high and low values of a VARIABLE. For example, suppose the variable is age, and the youngest person in the study is 26, the oldest 93. Then, the range of the age variable is 26 to 93, or 67 ($93 - 26 = 67$). The range is a quick first measure of the DISPERSION of a variable. However, it should not be used as the last measure of dispersion, because it is unduly influenced by OUTLIERS. In our example, suppose the next highest age after 93 was 61. The presence of the 93-year-old made the range great, perhaps suggesting there were many more very old people than there actually are.

—Michael S. Lewis-Beck

RANK ORDER

Rank order refers to a property of a set of scores such that the scores are ordered from lowest to highest, or, alternatively, from highest to lowest. The scores 15, 14, 8, 11, 12 can be placed into rank order by ordering them from lowest to highest as 8, 11, 12, 14, 15. These scores can be rank ordered from highest to lowest by putting them in the order 15, 14, 12, 11, 8. Stated more formally, a set of scores is rank ordered from lowest to highest if, for any given score X_j , it is always the case that $X_{j-1} \leq X_j \leq X_{j+1}$. Scores are rank ordered from highest to lowest if, for any given score X_j , it is always the case that $X_{j-1} \geq X_j \geq X_{j+1}$.

Sometimes, scores are assigned a rank based on their rank order. Ranks are consecutive, ordered numbers (usually integers) in which the number reflects the relative position of a score in a set of scores that is in rank order. For example, the scores 8, 11, 12, 14, and 15 might be assigned the following ranks:

Table 1

Score	Rank
15	5
14	4
12	3
11	2
8	1

A rank of 3 means that the score is the third highest score relative to the lowest score when scores are assigned ranks where the lowest score receives a rank of 1. A rank of 4 means the score is the fourth highest score relative to the lowest score. Sometimes, scores are assigned ranks whereby the highest score has a rank of 1 and the lowest score has a rank of N , where N is the number of scores being ranked.

The same score may appear twice in a set of scores, such as 8, 11, 15, 14, 8, 12. These scores can still be rank ordered from lowest to highest as 8, 8, 11, 12, 14, 15, or from highest to lowest as 15, 14, 12, 11, 8, 8. When assigning ranks to duplicate scores, the tradition is to assign a rank value that is the average of the ranks, as follows:

Table 2

Score	Rank
15	6
14	5
12	4
11	3
8	1.5
8	1.5

where the rank assigned to the two values of 8 is based on $(1 + 2)/2 = 1.5$.

—James J. Jaccard

RAPPORT

Rapport refers to the situation in interviewing when the interviewer and the interviewee feel comfortable and are in tune with each other. A rapport suggests each knows what the other is saying.

—Tim Futing Liao

RASCH MODEL

The Rasch model is a specific and structurally simple model in the family of ITEM RESPONSE models. The BINARY response [e.g., correct (score 1), incorrect (score 0)] to an item depends on a person's latent trait value and the difficulty of the item. The function that defines the relationship between a 1 score and the latent trait is logistic, with a unique location (difficulty) parameter for each item, but the same discrimination power.

—Klaas Sijtsma

See also ITEM RESPONSE THEORY

RATE STANDARDIZATION

Social scientists have been, more or less explicitly, comparing rates since they opened their eyes to the empirical world. But it wasn't until the late 1800s and early 1900s that scholars began to take seriously, and attempt to adjust for, the observed differences in rates that may be due to differences in the distribution(s) of some CONFOUNDING factor. For example, we may be interested in comparing some crime rate between two locales. If these two locales differ significantly in their age DISTRIBUTIONS, and if the crime rate in question is known to vary by age, then the crime rates will necessarily vary across the two locales to the extent that the age distributions are different. That is, any observed difference in the crime rates would necessarily be due in part to the differences in the age distributions. Moreover, that part of the difference in their rates could not be attributable to any difference in what may be considered the intrinsic nature of the locales facilitating or inhibiting crime. This, of course, has obvious consequences when informing any related social

policy. When comparing rates, therefore, an important question is, To what extent are observed differences in rates due to what are often called *compositional differences* across the populations being compared?

In the first half of the 20th century, part of the solution to this difficult question was to consider what the difference between two (or more) rates would have been if the populations being compared had the same composition (however defined, but frequently in terms of age, race, and/or sex distributions). To do this, researchers often would take as the standard compositional distribution(s) from one population, which could be one of the populations being compared or some other population external to the study. Rates for the remaining comparison population(s) were then calculated in such a way as to adjust, or control, for the compositional differences between the remaining population(s) and the standard population. These adjusted rates are referred to as *standardized rates*. Standardized rates are, to use current language, COUNTERFACTUAL. That is, they are not observed, but rather constructed so that they would have been observed had the compositional distributions been the same as the so-called standard. Thus, any difference in the standardized rates cannot be due to differences in compositional distributions across comparison populations.

Comparisons of the differences in the standardized (counterfactual) rates with the differences in the unstandardized (observed or crude) rates were used by researchers in the early part of the 20th century to calculate that portion of the observed difference due to compositional differences (see, e.g., Groves & Ogburn, 1928). But these applications remained rather informal until the late 1940s and early 1950s, when Evelyn Kitagawa and her colleagues at the University of Chicago formalized this comparison. This technique, published in 1955 by Kitagawa and referred to then as the "components of a difference between two rates" (Kitagawa, 1955, p. 1169), constituted a significant breakthrough and showed precisely how differences in observed crude rates were functions of, for example, compositional differences and composition-specific observed rates.

Since Kitagawa's (1955) fundamental publication, many scholars have added useful methodological advances to our understanding of rate standardizations and comparisons. Among the most important is Clogg's (1978) use of LOG-LINEAR MODELS to develop general methods for what he termed *purging rates* of confounding effects. Clogg's (1978) purging

method—along with elaborations and extensions developed in Clogg and Eliason (1988); Liao (1989); Xie (1989); and Clogg, Shockey, and Eliason (1990)—provides a powerful modeling approach, with foundations in the apparatus of statistical inference, to compare observed, standardized, and, more generally, purged rates, as well as functions of these rates.

—Scott R. Eliason

REFERENCES

- Clogg, C. C. (1978). Adjustment of rates using multiplicative models. *Demography*, 15, 523–539.
- Clogg, C. C., & Eliason, S. R. (1988). A flexible procedure for adjusting rates and proportions, including statistical methods for group comparisons. *American Sociological Review*, 53, 267–283.
- Clogg, C. C., Shockey, J. W., & Eliason, S. R. (1990). A general statistical framework for adjustment of rates. *Sociological Methods & Research*, 19, 156–195.
- Groves, E. R., & Ogburn, W. F. (1928). *American marriage and family relationships*. New York: Holt.
- Kitagawa, E. (1955). Components of a difference between two rates. *Journal of the American Statistical Association*, 50, 1168–1194.
- Liao, T. F. (1989). A flexible approach for the decomposition of rate differences. *Demography*, 26, 717–726.
- Xie, Y. (1989). An alternative purging method: Controlling the composition-dependent interaction in an analysis of rates. *Demography*, 26, 711–716.

RATIO SCALE

S. S. Stevens (1946) described four levels of measurement ranging from least to most precise: nominal, ordinal, INTERVAL, and ratio. Nominal, ordinal, and interval scales are described elsewhere in this volume.

Ratio scales share several common characteristics with interval scales. They distinguish greater and lesser amounts. Intervals between any two adjacent numerals indicating quantities are equal in size.

The distinguishing characteristic of ratio scale is the existence of an absolute origin or zero point. This zero allows the researcher to make explicit how much greater or smaller one quantity is than another. Such statements are not possible with interval scales. [Others argue that the distinction between interval and ratio measurement is of no consequence (Wright, 1997).]

For example, my weight is 180 pounds. My wife weighs 120. We could say, in comparing the two of us, that I am the heavier individual, an ordinal level

distinction. Or, it is possible to state that I weigh 60 pounds more than my wife, a result of the interval-level measurement of weight. Or, I could indicate that weight is a ratio variable by asserting that I am 1.5 times heavier than my wife (and twice as heavy as my 90-pound son). On the other hand, if we were to compare today's cold 30-degree temperature to yesterday's warm 60-degree day, it is possible to claim it is 30 degrees colder today (if we're measuring temperature in Fahrenheit degrees), but we certainly couldn't claim that it is half as warm.

A more subtle way of distinguishing the levels of measurement is to identify the nature of invariance implied by each. Ordinal scales may be subject to MONOTONIC transformations and retain the information in their original scale. Interval scales may be LINEARLY TRANSFORMED. Ratio scales may be transformed only by a CONSTANT. That is, they retain their original information only when subject to multiplication by a constant term ($X' = aX$). Only under this unique situation do ratios remain the same. If I gained 180 pounds (doubling my weight) and my wife gained 120 pounds (getting heavier by a factor of 2), we would retain the same ratio of weight (1.5).

Conventional wisdom has accepted Stevens's argument that certain statistical procedures are appropriate for data at each LEVEL OF MEASUREMENT. For example, the only measure of CENTRAL TENDENCY appropriate for nominal data is the MODE, whereas MEDIANS can be used to describe ordinal data, and arithmetic MEANS are applicable if the data are measured on interval scales.

Ratio-scaled data seem intrinsically valuable, but in many instances, social scientists do not aspire to such precision. Exploratory studies often focus on nominal and ordinal levels of measurement of phenomena of interest. Most CORRELATIONAL/causative studies seek interval-level data. The most common assertion for ratio data comes with the examination of models with INTERACTION terms. A natural origin provides for a fixed interpretation of these terms not always possible with interval-level data centered at arbitrary zero points.

Despite apparent common acceptance of Stevens' typology of levels of measurement—at least as a didactic tool for teaching undergraduates—considerable disagreement surrounds his prescription that level of measurement proscribes the use of certain statistics (Baker, Hardyck, & Petrinovich, 1966; Borgatta & Bohrnstedt, 1980; Gaito, 1980; Mitchell, 1986; Townsend & Ashby, 1984; Velleman & Wilkinson, 1993).

A middle ground might be to recognize that level of measurement is not an innate characteristic of the object being measured. That is, the same measurement of an object may well be treated as representing any of the levels of measurement identified by Stevens. Similarly, once we recognize the likelihood of ERROR—SYSTEMATIC or RANDOM—in any measurement, we can find ourselves less likely to assert strict adherence to a specific level of measurement as a defining characteristic of an object. The measuring instrument also plays a role in establishing the level of measurement of any phenomenon under study.

Stevens' taxonomy of measurement, although the most common, is not the only existing classification scheme. Mosteller and Tukey (1977) and Chrisman (1997), for example, offer a number of additions and modifications to Stevens' taxonomy. Most important about these alternatives is not that they are necessarily better or worse, but rather that they make it clear that Stevens' categorization is not exhaustive.

—John P. McIver

REFERENCES

- Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurement vs. strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, 26, 291–309.
- Borgatta, E. F., & Bohrnstedt, G. W. (1980). Level of measurement—Once over again. *Sociological Methods & Research*, 9, 147–160.
- Chrisman, N. (1997). *Exploring geographic information systems*. New York: Wiley.
- Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 87, 564–567.
- Koch, W., Schulz, E. M., Wright, R., Smith, R. M., Lang, S. (1996). What is a ratio scale? *Rasch Measurement Transactions*, 9, 457.
- Mitchell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, 100, 398–407.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.
- Stevens, S. S. (1951). *Handbook of experimental psychology*. New York: Wiley.
- Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin*, 96, 394–401.
- Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *American Statistician*, 47, 65–72.
- Wright, B. D. (1997). S. S. Stevens revisited. *Rasch Measurement Transactions*, 11, 552–553.

RATIONALISM

This term refers both to a general philosophical position and to a group of theories of human action and social relations. The long-standing opposition in Western philosophy between rationalism and EMPIRICISM was over whether our knowledge originates in reason or experience, or which of them has priority. In modern philosophy of science, rationalists claim that laws of science are necessarily true, and this idea of necessity sometimes brings them close to REALISM and its account of CAUSALITY. Rationalism is, however, used more often to refer to the principle that humans mostly act or can be assumed to act for good reasons, and that social explanations should be in terms of (usually individual) rational action. This generates a powerful research program in economics and the other social sciences: If we know that it is rational to sell something for the highest price, we can predict that, other things being equal, people will do this. As the philosopher Martin Hollis (1977) put it, “rational action is its own explanation” (p. 21): It is only when people fail to act rationally that we require a special explanation of their conduct. There are, of course, different levels of rational explanation: It is rational to buy cigarettes at the cheapest price, but irrational to smoke them at all. Rational action and rational choice are not necessarily self-interested; people may rationally choose to travel by train rather than car or plane because it causes less environmental damage.

There is substantial disagreement over the usefulness of rational action and rational choice explanations, although they have become increasingly prominent in political science and international relations and can even be found in sociology and social anthropology. Are we really as calculating in our choices of, say, marital partners, political parties, or religions as the model suggests? Rational action theories may seem to ignore the rich texture of human cultures and ideologies. On the other hand, they have given social science some valuable tools and explanatory paradigms, such as the “Free Rider,” the PRISONER’S DILEMMA, and

other models from GAME THEORY. More generally, an understanding (see VERSTEHEN) of the ways in which people understand their situations and the reasons why they act as they do is a central element in social inquiry, even if it is ultimately supplemented or sidelined by more structural explanations. It is worth remembering that we are most often asked to “explain ourselves” when we are thought to have done something wrong.

—William Outhwaite

REFERENCES

- Elster, J. (Ed.). (1986). *Rational choice*. Oxford, UK: Basil Blackwell.
- Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge, UK: Cambridge University Press.
- Hollis, M. (1977). *Models of man*. Cambridge, UK: Cambridge University Press.
- Hollis, M., & Smith, S. (1990). *Explaining and understanding international relations*. New York: Oxford University Press.

RAW DATA

Often, we lament, “So much to do, so little time.” The same complaint applies to DATA. Frequently, we find ourselves with “so much (raw) data, and so little (useful) information.” The challenge for most researchers who collect and analyze data is to extract useful information from the raw materials with which they start.

Let’s begin with a series of distinctions necessary to describe raw data.

First of all, “data” is a plural noun. Data are multiples of individual datum.

Second, all collections of data are incomplete. Data are uncollected because of a variety of constraints. Time, money, and their combination, efficiency, are three common reasons we fail to collect all available raw data to answer any question. Increasingly, however, data collection has become easier and more efficient, and more and more data are collected. Absent ability to process this additional raw data, more data may not increase knowledge, but enhanced computing power has produced an ability to process more data more efficiently.

Third, the expression “raw data” implies a level of processing that does not meet expectations that these data provide useful information without further effort. “Raw” suggests that the data have not yet been

summarized or analyzed in a way as to release the information for which they were collected.

Fourth, raw data may be collected in alternative representational schemes that affect how the researcher thinks about processing them. Perhaps the easiest way to think about this is to consider the variety of symbolic formats used in computer processing of information. Data are commonly represented in binary, hexadecimal, or decimal number systems, or as (ASCII) text. Data organization is also relevant: Information may be organized as bits, nibbles, bytes, words, and so on.

One individual’s processed data are another’s raw data. The methods by which data are collected and the level of aggregation of the collected data play important roles in determining how the material gathered as raw data is to be analyzed.

Consider, for example, the collection of responses to a series of five SURVEY questions designed to measure public opinion about abortion. The most disaggregated version of these data is the individual responses—typically represented as numbers symbolizing the answers—to each of the five questions. These data might be used by a survey researcher to establish the patterns or consistency of responses of individuals to the survey questions. Such efforts might produce a second level of data in which a SCALE is created to represent the set of responses to the series of questions, with one score assigned to each individual. A third level of data aggregation might be reached by summarizing the individual scale scores for subsamples or groups of the individuals interviewed. A fourth level of aggregation of these data would be to report what Page and Shapiro (1992) call “collective public opinion” and report the average scale score for the entire SAMPLE—usually in comparison to another sample taken from the same POPULATION at a different point in time or to a sample drawn from a different population at the same point in time.

Even more critical than the methodology used to collect data is the RESEARCH QUESTION to be answered or the THEORY to be tested. Raw data must be more (actually less) than simply the set of all information available in the world. What distinguishes raw data from grains of sand? Data must be identified and collected with a purpose. Research questions motivate the collection of some subset of all possible data and the exclusion of other information.

Ultimately, all data should be subject to a variety of assessments, including judgment as to the degree of error contained in the data relevant to the purpose for

which they were collected. Such assessment includes establishing the RELIABILITY and VALIDITY of data as well as justifications for the discarding of data and consideration of the problem of MISSING DATA.

Other references to data described in the encyclopedia include ARCHIVAL RESEARCH, AUTOBIOGRAPHY, CATEGORICAL DATA, CENSORED DATA, COMPUTER-ASSISTED DATA COLLECTION, CROSS-SECTIONAL DATA, DATA, EXPERIMENTS, INTERNET SURVEYS, ORAL HISTORY, PANEL DATA ANALYSIS, PERSONAL DOCUMENTS, SECONDARY DATA, and TIME SERIES.

—John P. McIver

REFERENCES

- Coombs, C. H. (1964). *A theory of data*. New York: Wiley.
- Diesing, P. (1991). *How does social science work?* Pittsburgh, PA: University of Pittsburgh Press.
- Hyde, R. (2003). *The art of assembly language*. San Francisco: No Starch Press.
- Page, B. I., & Shapiro, R. Y. (1992). *The rational public*. Chicago: University of Chicago Press.

RC(M) MODEL. See ASSOCIATION MODEL

REACTIVE MEASURES. See REACTIVITY

REACTIVITY

The term *reactivity* simply means that measurements or observations of behavior may influence the behavior, thereby making interpretations subject to error. The reactivity problem can thus be understood as one facet of the wider problem of research artifacts (Rosnow & Rosenthal, 1997), which may be manifested in different ways. For example, based on analyses of filmed interactions between experimenters and research subjects, Robert Rosenthal (1966) observed that experimenters smiled more at female than male subjects, which he whimsically interpreted as suggesting that “chivalry is not dead in the psychological experiment.” Smiling by experimenters affected the subjects’ responses, which is interesting psychologically but disconcerting methodologically,

Rosenthal cautioned. He attributed biasing effects of experimenters to uncontrolled psychosocial and biosocial factors, but he pointed out that experiments with nonhuman subjects are susceptible to reactivity effects as well. For example, he mentioned the finding that experienced observers in an animal laboratory were able to guess which of several experimenters had been handling a rat from the animal’s behavior while running a maze or when being picked up.

In research in which self-report personality measures are used, the reactivity problem is sometimes manifested as SOCIAL DESIRABILITY BIAS, meaning that respondents are reluctant to reveal their true feelings or beliefs, but instead respond in ways they think will elicit favorable impressions of themselves. One standard method for identifying this type of bias is to correlate responses on the personality measures and a measure of social desirability. There is now a body of research and theory on self-report measures, including procedures for detecting and, sometimes, controlling reactivity (Stone et al., 2000).

Another distinction is that made between *reactive measures* and *nonreactive measures*, the latter referring to measures that do not influence behavior (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). Among the various types of nonreactive measures used by social researchers are physical traces, covert instruments, and contrived unobtrusive observations. For instance, Webb et al. cited researchers who had used pictures of art objects depicting how infants and small children were portrayed to study sociohistorical similarities and differences. The use of DECEPTION (e.g., concealed microphones or hidden cameras) frequently raises ethical concerns, but Webb et al. argued that, in most instances of social research, the covert measures used are harmless enough so that they do not raise serious ethical problems. Research employing contrived, unobtrusive observation calls not only for the use of hidden measures, but for some type of manipulation, as in certain FIELD EXPERIMENTS in the social sciences.

—Ralph L. Rosnow

REFERENCES

- Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.
- Rosnow, R. L., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. New York: W. H. Freeman.

- Stone, A. A., Turkkan, J. S., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (Eds.). (2000). *The science of self-report: Implications for research and practice*. Mahwah, NJ: Lawrence Erlbaum.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. B. (1981). *Nonreactive measures in the social sciences* (2nd ed.). Boston: Houghton Mifflin.

REALISM

Realists assert the actual or possible existence of some disputed entity, which may be a subatomic particle, a social structure, or the world or universe itself. Realism as a general philosophical position is often opposed to skepticism, the doctrine that we can never know whether something (or even anything) exists, and to positions that see reference to entities as merely a useful fiction to make sense of observations. Realism tends to be associated with materialism as opposed to IDEALISM, although it is opposed to nominalism about “universals” or the referents of general terms. In this latter sense, it is closer to idealism in asserting the real existence of, say, something called redness as distinct from its instantiation in particular red objects. In modern philosophy of science, realism mainly opposed POSITIVISM and the claim made in logical positivism or logical empiricism that metaphysical questions about the reality of objects, as distinct from our observations of them, are meaningless. More recently, with the eclipse of positivism (see POSTEMPIRICISM), realists, especially those identifying with CRITICAL REALISM, have turned their opposition more to social CONSTRUCTIONISM and POSTMODERNISM. Others, however, combine a realist conception of science with a social constructionist account of human societies.

Arguments for realism fall into two broad classes. First, there are inductive arguments from scientific convergence. Where scientists increasingly agree about explanations that have been refined over many years, it may be reasonable to think that we are getting closer to a true representation of reality. Second, there are arguments of principle based on what is required for science to be possible or meaningful—notably, that the world exists in an ordered and predictable way. In the social sciences, both sets of arguments are more problematic: There is less convergence on agreed PARADIGMS and not even complete agreement on the very possibility of social science (see NATURALISM). Social scientists are more likely to precede their references to social

structures and mechanisms by an implicit “It is as if . . .”

Realism does not entail a commitment to any particular theory of the social world, nor to any particular set of research methods. As an anti- or postempiricist position, however, it plays down the importance of repeated observation and prediction as criteria of explanatory success. Entities that are observable only in their effects, such as magnetic fields and social structures, may interact with others in such a way as to neutralize one another, although theoretical analysis and experimentation may, nevertheless, demonstrate their existence. Here on the Earth’s surface, we are subject both to its gravitational attraction and to the centrifugal force of its rotation; the result is that we mostly stay where we are. Social forces, individual or structural, may interact in a similar way, but again, they are harder to demonstrate to general satisfaction. Antirealists are suspicious of an apparently uncontrolled proliferation of entities, endowed like humans with causal powers (see CAUSALITY), and accuse realists of ESSENTIALISM.

—William Outhwaite

REFERENCES

- Outhwaite, W. (1987). *New philosophies of social science: Realism, hermeneutics and critical theory*. Basingstoke, UK: Macmillan.
- Sayer, A. (1992). *Method in social science: A realist approach* (2nd ed.). London: Routledge.
- Sayer, A. (2000). *Realism and social science*. London: Sage.

RECIPROCAL RELATIONSHIP

Reciprocal relationship refers to a hypothesized set of relationships in which some event *X* causes a second event *Y*, and vice versa. Several statistical techniques are used to determine whether this is the case, including the *F* RATIO, indirect least squares (ILS), the INSTRUMENTAL VARIABLES approach, three-stage least squares (3SLS), and TWO-STAGE LEAST SQUARES (also known as nonrecursive causal modeling). Reciprocal relationships are common in CAUSAL MODELING, or in structural equation estimation.

Consider the case of two-stage least squares (2SLS), which is a special case of the instrumental variables approach (Kennedy, 1998, p. 165). To perform 2SLS, one must estimate two equations, one after the other. The first equation should “regress each

ENDOGENOUS VARIABLE acting as a regressor in the equation on all the EXOGENOUS VARIABLES in the system of simultaneous equations ... and calculate the estimated values of these endogenous variables” (Kennedy, 1998, p. 165). The second equation should “use these estimated values as instrumental variables for these endogenous variables or simply use these estimated values and the included exogenous variables as regressors in an OLS [ORDINARY LEAST SQUARES] equation” (Kennedy, 1998, p. 165). This approach tests for the possibility of a reciprocal relationship as it individually regresses all endogenous variables on all of the exogenous variables in that system of simultaneous equations and then calculates estimated values to be used as regressors in ordinary least squares.

An illustration of a reciprocal relationship is useful to fully understand this concept. Political scientist B. Dan Wood illustrated how a reciprocal relationship works in his study on environmental politics. Wood argued that federal implementation of environmental laws in the United States that pertain to clean air is dependent upon both national and subnational implementation of these laws (Wood, 1992). He tested this argument via an instrumental variables approach in which he designed his first equation to predict state-level implementation. His second equation took the predicted values from the first equation and used them as an additional INDEPENDENT VARIABLE to predict federal implementation of environmental laws. A reciprocal relationship exists in this case because the level of state enforcements may depend on the fiscal support that the United States’ Environmental Protection Agency provides to state programs, but state activities may also determine the level of these supports (Wood, 1992, p. 50). This reciprocal relationship is important because Wood (1992) had to determine two things to ensure that this reciprocal relationship exists: (a) the effect of the level of federal fiscal support on the level of state enforcements, and (b) the effect of the level of state enforcements on how much federal fiscal support was received by an individual state. Wood (1992) applied an instrumental variables approach and used his results to conclude that a reciprocal relationship exists between the level of state enforcements and fiscal support provided by the Environmental Protection Agency.

—Kenneth W. Moffett

REFERENCES

- Kennedy, P. (1998). *A guide to econometrics* (4th ed.). Cambridge: MIT Press.
- Kmenta, J. (1997). *Elements of econometrics* (2nd ed.). Ann Arbor: University of Michigan Press.
- Wood, B. D. (1992). Modeling federal implementation as a system: The clean air case. *American Journal of Political Science*, 36, 40–67.

RECODE

Recode is a term used to describe the process of making changes to the values of a VARIABLE. Rarely can researchers proceed directly to data analysis after receiving or compiling a data set. The values of the variables to be used in the analysis usually have to be changed first. The reasons for recoding a variable are many, and include the following:

1. There may be errors in the original CODING of the data that must be corrected. A FREQUENCY DISTRIBUTION run or descriptive statistics on a variable can help the researcher identify possible errors that must be corrected with recodes.

2. Data are usually collected to get as much information as possible, but these data often must be recoded to yield more interpretable results. For example, respondents may be asked to report their date of birth in a survey. Recoding these values to age (in years) lets the researcher interpret the variable in a more intuitive way.

3. Recoding is also necessary if some values of a variable are coded inconveniently. For example, surveys typically code responses of “Don’t know” and “Refused” as 8 and 9. The researcher needs to recode these values so they are recognized as missing values by the computer program being used. Leaving these values unchanged would yield misleading results.

4. Another common reason to recode a variable is to make it ordinal in its codes. The respondent’s party identification may be coded 1 = Democrat, 3 = Republican, and 5 = Independent in a survey. This can be recoded as 1 = Democrat, 2 = Independent, and 3 = Republican to make it ordinal.

5. Recoding can also be used to collapse the values of a variable into fewer categories. For example, respondent’s income may have been collected in \$10,000 ranges where 1 = \$0 to \$10,000, 2 = \$10,001 to

\$20,000, 3 = \$20,001 to \$30,000, and so on, for a total of 10 categories. Income can now be recoded into low, medium, and high levels of income, where the old values of 1, 2, and 3 are recoded to the new value of 1 (low income) and the old values of 4, 5, and 6 are recoded to the new value of 2 (medium income), and so on.

6. Another reason to recode a variable is to transform the values of a variable (e.g., a log transformation) to tease out the true nature of the relationship between two variables.

7. Also, if missing values are assigned a number in the original coding, the researcher needs to recode these values to missing before analyzing the data. It is advisable to run a frequency or descriptive statistics on the recoded variable to make sure that the desired recodes were achieved.

—Charles Tien

See also CODING

REFERENCES

- Nagler, J. (1995). Coding style and good computing practices. *Political Methodologist*, 6, 2–8.
- Norusis, M. J. (1990). *SPSS base system user's guide*. Chicago: SPSS.

RECORD-CHECK STUDIES

Record-check studies obtain the same factual data from two sources, typically a household survey and administrative records of some kind. Most often, the objective is assessment of measurement error in the household survey; such cases are often called validation studies, and the records data are typically presumed to be without error. Occasionally, data from records may replace that reported in the household survey in calculating survey estimates. As described by Machlin and Taylor (2000), the Medical Provider Survey component of the Medical Expenditure Panel Survey is an example of a data replacement record check.

Marquis (1984) and Groves (1989) each provide detailed definitions of validation-type record-check studies. Marquis described a basic typology for record checks in terms of the values obtained for a binary

variable (e.g., the presence or absence of a chronic condition) in a two-by-two matrix. The two sources could agree positively on the presence of the condition (Cell A of the matrix), they could agree negatively (Cell D), or either the records (Cell B) or the survey (Cell C) could be positive and the other negative. A full-design record check would allow quantification of all four cells. In some “reverse record check” studies, records are checked only for survey respondents reporting positively (an “AC” design). In some “forward record check” studies, a survey is conducted only with individuals identified positively in the records (an “AB” design).

An example of a full-design record-check study is described by Edwards et al. (1994), who compared reports of chronic conditions in a study using the U.S. National Household Interview Survey (NHIS) questionnaire with respondents selected from the membership of a staff model health maintenance organization (HMO). They reported on the agreement between the two sources using the kappa statistic and found good to excellent agreement on some conditions (e.g., diabetes and hypertension) but poor agreement on others (e.g., chronic bronchitis, heart murmurs). They also compared the population estimates resulting from each source and found both overreporting in the household survey as compared with the records (e.g., heart murmurs, tinnitus) and underreporting (e.g., dermatitis, heart conditions).

Two limitations of validation-type record checks identified by Groves are (a) that the records are presumed to be error-free, when, in fact, record systems are often subject to the same kinds of errors as surveys, and (b) that household reports and records data must be matched before they may be compared, another process subject to ambiguity and error. In the Edwards et al. study, for example, a “true” medical condition may have been missing from the records because the patient never mentioned it, because it was diagnosed and treated before the patient joined the HMO, or because of error by medical personnel. Matching medical events, as described by Machlin and Taylor for the MEPS, is a probabilistic process because of errors in the matching characteristics, such as name of patient, data of event, type of event, and reason for the event.

—W. Sherman Edwards

REFERENCES

- Edwards, W. S., Winn, D. M., Kurlantzick, V. et al. (1994). *Evaluation of National Health Interview Survey diagnostic reporting*. National Center for Health Statistics. Vital Health Stat 2(120).
- Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
- Machlin, S. R., & Taylor, A. K. (2000). *Design, methods, and field results of the 1996 Medical Expenditure Panel Survey Medical Provider Component* (MEPS Methodology Report No. 9, AHRQ Pub. No. 00-0028). Rockville, MD: Agency for Healthcare Research and Quality.
- Marquis, K. (1984). Record checks for sample surveys. In T. Jabine, E. Loftus, M. Straf et al. (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines*. Washington, DC: National Academy Press.

RECURSIVE

Many social scientific THEORIES specify models of more than one equation. A recursive system of equations is a type of SIMULTANEOUS EQUATION system where the equations are linked through a causal chain of dependent variables. In recursive models, (a) direction of influence in the model moves only one way (there is no reciprocal causation or feedback loops), and (b) the errors in the equations are not correlated. Both conditions must be met for a system of equations to be considered recursive.

Consider the following recursive system of equations:

$$\begin{aligned}y_1 &= \gamma_{11}x_1 + u_1, \\y_2 &= \beta_{21}y_1 + \gamma_{21}x_1 + \gamma_{22}x_2 + u_2, \\y_3 &= \beta_{31}y_1 + \beta_{32}y_2 + \gamma_{32}x_2 + u_3,\end{aligned}$$

where $\text{COV}(u_1, u_2) = \text{COV}(u_1, u_3) = \text{COV}(u_2, u_3) = 0$.

In the system of equations, ENDOGENOUS VARIABLES are denoted with y s, and EXOGENOUS VARIABLES are denoted with x s. β s indicate the effect of an endogenous variable on another endogenous variable, and γ s indicate the effect of an exogenous variable on an endogenous variable. Errors in the equations are indicated by u s.

The model can be drawn as a PATH DIAGRAM, which visually presents the relationships between the variables (see Figure 1).

The MODEL can also be represented as MATRICES of coefficients. Note that the subscripts in the equations

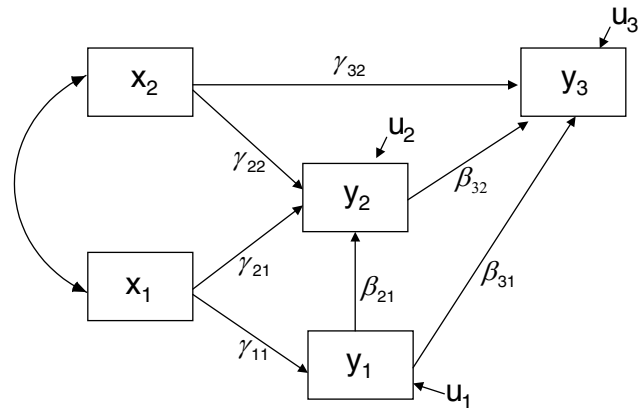


Figure 1

above denote the position of the coefficient in the matrix.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} \gamma_{11} & 0 \\ \gamma_{21} & \gamma_{22} \\ 0 & \gamma_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}.$$

As in all recursive systems of equations, the above model can be viewed in causal sequence (y_1 affects y_2 , y_1 and y_2 affect y_3), and the disturbances for one equation are uncorrelated with the disturbances from the other equations. Formally, in recursive models, **B** may be written as a lower triangular matrix (coefficients appear only below the diagonal) and the variance-covariance matrix of the errors in the equations (u s) is a diagonal matrix. If there was any reverse causality in the model, such as an effect of y_3 on y_1 (either directly or through a chain of other variables), or if any or all of the errors were correlated, then the model would be NONRECURSIVE.

Each equation of a recursive model can be estimated separately using ORDINARY LEAST SQUARES (OLS) regression.

—Pamela Paxton

See also STRUCTURAL EQUATION MODELING

REFERENCES

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Greene, W. H. (1997). *Econometric analysis* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Heise, D. R. (1975). *Causal analysis*. New York: Wiley.

REDUCTIONISM

In philosophy of science, the term *reductionism* is used in a general sense to indicate the view that complex explanations can or should be reduced to simpler ones—for example, that biological phenomena can be explained by their chemical constituents, which in turn can be explained by their atomic or subatomic properties. In social science, reductionism operates at different levels but follows the same principle of seeking an explanation of complex phenomena in terms of simpler ones. It is a controversial strategy, and those who pursue it rarely use the term, preferring (for example) descriptions such as *METHODOLOGICAL INDIVIDUALISM*.

Reductionism is usually a more radical strategy than individualism and can apply to a wide range of phenomena. It has been pursued in most social sciences, particularly economics, politics, anthropology, and sociology. The approach might be illustrated by the contemporary revival of the Victorian passion for reducing explanations of social phenomena to individual, inherited biological characteristics. This has found its modern expression in the discipline of evolutionary psychology. The extent to which the biological can be said to determine the social, or the form it takes, differs widely among the proponents of an evolutionary approach, but the most common form taken is that of explanations that depend on genetic inheritance.

There are two components to this approach—the genetic one, which moves beyond the more established view that our genetic inheritance can predispose us to certain biological traits (diseases such as cancer, for example), to one that says our genes are wholly or partly responsible for psychological characteristics such as intelligence, or particular forms of behavior such as deviance, spatial ability, sociability, and so on. Second, evolutionary psychologists maintain that genetic makeup results from natural selection in the human species. The Darwinian principle of natural selection rules out the transmission of learned characteristics to offspring; therefore, only reproductive transmission can pass on the genes that lead to biological and behavioral traits. Humans have changed little in physiological terms since the beginning of civilization, specifically in brain capacity and characteristics. Therefore, it is said to follow that our contemporary genetic endowment was shaped at a time when our brains were evolving differently from our

primate relatives. The explanation does not rule out environmental interaction, but it is argued that if we want an explanation for human behavior, we must look to the genetic inheritance of behavioral traits first established precivilization.

Reductionist explanations such as these have led to some very controversial social research. Most notorious is Herrnstein and Murray's *The Bell Curve* (1994). Their central argument was that the majority of America's social and economic problems are the result of low intelligence among those who suffer poverty, unemployment, and educational failure. Thus, it follows that no amount of social intervention—HeadStart programs, for example—can make much difference to the life chances of the individuals and, by extension, the social milieu in which they live. Their argument rests on an analysis of IQ data, from the National Longitudinal Survey of Youth, that they claim demonstrated that IQ tests were a reliable predictor of life success (Dickens, 2000, p. 67). Their research has been criticized both for the weakness of the correlations that led to these conclusions and for the normatively derived definition of IQ.

More generally, critics of biological reductionist approaches focus on the weakness of the data supporting the arguments and the tendency to introduce auxiliary hypotheses to explain empirical contradictions (Dupré, 2001, p. 68). However, the challenge for those in the social sciences, who dismiss *any* biological antecedents for behavior, is to explain how social behavior developed independently of biology.

—Malcolm Williams

REFERENCES

- Dickens, P. (2000). *Social Darwinism*. Buckingham, UK: Open University Press.
- Dupré, J. (2001). *Human nature and the limits of science*. Oxford, UK: Oxford University Press.
- Herrnstein, R., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

REFLEXIVITY

Reflexivity is a term that is used in a variety of ways (see Bourdieu & Wacziarg, 1992; Lynch, 2000). For some writers, such as Mead and Garfinkel, reflexivity is a central feature of human social life, although what

they mean by the term differs. For Mead, an individual becomes a self by being able to take the attitude of others and thereby reflect on his or her own behavior (see SYMBOLIC INTERACTIONISM). For Garfinkel, reflexivity refers to the fact that social actions are essentially accountable, in the sense that they are what they are only in and through what they are displayed and read as being by members (see ETHNOMETHODOLOGY). Other usage of reflexivity, and the main focus here, refers to a feature of social research practice, both an existential fact about it and something that can and should be made a virtue. It is argued that researchers are always part of the social world they study; they can never step above it in order to gain an Olympian perspective or move outside it to get a “view from nowhere.” It is taken to follow from this that they should continually reflect on their own role in the research process and on the wider context in which it occurs. This kind of reflection should guide the course of research; and in writing up their findings, it is argued, researchers ought to give a detailed account of the research process so as to allow readers to judge their findings in context (e.g., see Hammersley & Atkinson, 1995).

Several arguments are offered in favor of researcher reflexivity, some of which are incompatible with one another. The most fundamental and straightforward is that research is a complex and difficult activity that cannot be reduced to following a set of preestablished rules. Rather, deciding how best to do it requires continual reflection on what has been done, how successful it has been, and how best to pursue it further. This is a relatively uncontroversial idea. A more developed version of this argument is that the researcher must reflect on how he or she has influenced the situation and the people being studied in order to monitor REACTIVITY, so as to minimize any distorting effect on the research findings.

Another argument for reflexivity demands that sociological theory be applied to social research itself. An example would be Becker’s application of labeling theory to the issue of research BIAS (Becker, 1967). A rather different one is Gouldner’s “reflexive sociology,” where sociology is seen as a transformative social movement that is continually building, applying, and developing an understanding of the world (and of itself) that can facilitate the realization of collective human goals (see Gouldner, 1973).

Other interpretations of reflexivity are linked with skeptical EPISTEMOLOGICAL ideas. The starting point

here is an insistence that knowledge is never simply a representation of what exists but always amounts to a social construction, and therefore reflects the perspective and social location of the researcher. Some versions of this argument are committed to the idea that through reflexivity, research can be made a transparent activity that generates genuine knowledge (see Alvesson & Skoldberg, 2000). For others, the virtue of reflexivity lies entirely in its continual subversion of any tendency to believe that accounts can represent the world. Here, reflexivity amounts to an ethical or political commitment to continually remind readers that all accounts, even including those of the constructionist researcher, are constructions rather than representations (see Woolgar, 1988). From this point of view, accounts cannot be judged in terms of VALIDITY, at least not conceived as correspondence to reality, but must instead be assessed according to political, ethical, or aesthetic criteria. In fact, the primary criterion may simply be how reflexive the researcher has been.

On the basis of all these arguments, reflexivity, albeit conceived in different ways, becomes the key virtue in research and, in some cases, the only possible one. However, there has been criticism of this idea, and there have also been disputes among those advocating different interpretations of reflexivity, or even purportedly operating within the same approach (see Czyzewski, 1994). Some criticism has come from both EMPIRICIST and antiempiricist directions, such as ethnomethodology (see Lynch, 2000). On one side, it is argued that reflecting on the course of the research process does nothing, in itself, to increase the likely validity of the findings. On the other side, some ethnomethodologists criticize this notion of reflexivity for itself implying that researchers can somehow rise above the reflexively constituted world of which both they and what they are studying are necessarily part.

Other criticism has rejected the idea that reflexivity can provide a basis for some single, comprehensive understanding of the social world in the way that, for example, Gouldner proposes (see Hammersley, 1999). Concerns have also been expressed about the consequences of requiring that researchers be reflexively accountable. In particular, it has been suggested that this may expose junior researchers to surveillance that could amount to a form of domination (Paechter, 1996; Troyna, 1994).

The term *reflexivity* is open to a variety of interpretations, then, and although particular versions have been strongly advocated by some as an important part of

postpositivist social research methodology, their value has been challenged by others.

—Martyn Hammersley

REFERENCES

- Alvesson, M., & Skoldberg, K. (2000). *Reflexive methodology*. London: Sage.
- Becker, H. S. (1967). Whose side are we on? *Social Problems*, 14, 239–247.
- Bourdieu, P., & Wacquant, L. J. D. (1992). *An invitation to reflexive sociology*. Chicago: University of Chicago Press.
- Czyzewski, M. (1994). Reflexivity of actors versus reflexivity of accounts. *Theory, Culture and Society*, 11, 161–168.
- Gouldner, A. V. (1973). *For sociology*. Harmondsworth, UK: Penguin.
- Hammersley, M. (1999). Sociology, what's it for? A critique of the grand conception. *Sociological Research Online*, 4(3). Retrieved from <http://www.socresonline.org.uk/socresonline/4/3/hammersley.html>
- Hammersley, M., & Atkinson, P. (1995). *Ethnography: Principles in practice*. London: Routledge.
- Lynch, M. (2000). Against reflexivity as an academic virtue and source of privileged knowledge. *Theory, Culture and Society*, 17(3), 26–54.
- Paechter, C. (1996). Power, knowledge and the confessional in qualitative research. *Discourse: Studies in the Politics of Education*, 17(1), 75–84.
- Troyna, B. (1994). Reforms, research and being reflexive about being reflective. In D. Halpin & B. Troyna (Eds.), *Researching education policy*. London: Falmer.
- Woolgar, S. (Ed.). (1988). *Knowledge and reflexivity*. London: Sage.

REGRESSION

Regression is the most widely used data analysis technique in the nonexperimental social sciences. In a regression model, a DEPENDENT VARIABLE, commonly labeled Y , is a function of one or more INDEPENDENT VARIABLES, commonly labeled X_1 , X_2 , and so on. The independent variables are assumed to explain, or at least predict, the phenomenon measured by the dependent variable. The simplest of regression models posits only one independent variable: $Y = a + b_1X_1 + e$. This bivariate regression says variable Y is a linear additive function of variable X_1 , plus a constant (the fixed value represented by a) plus error (represented by e). Of principal interest is the influence of X_1 on Y , the regression coefficient represented by b_1 . To obtain numerical estimates for the values a and b_1 , a straight line is fitted to the observations on X_1 and Y by the

method of ORDINARY LEAST SQUARES (OLS). A more complicated regression model posits two independent variables: $Y = a + b_1X_1 + b_2X_2 + e$. This multiple regression model says Y is a linear function of X_1 and X_2 (plus a constant term and an error term). A MULTIPLE REGRESSION ANALYSIS equation always has two or more independent variables. Below, the mechanics of bivariate and multivariate regression are reviewed in a DATA example. Then, assumptions, history, and recent developments are considered.

BIVARIATE REGRESSION

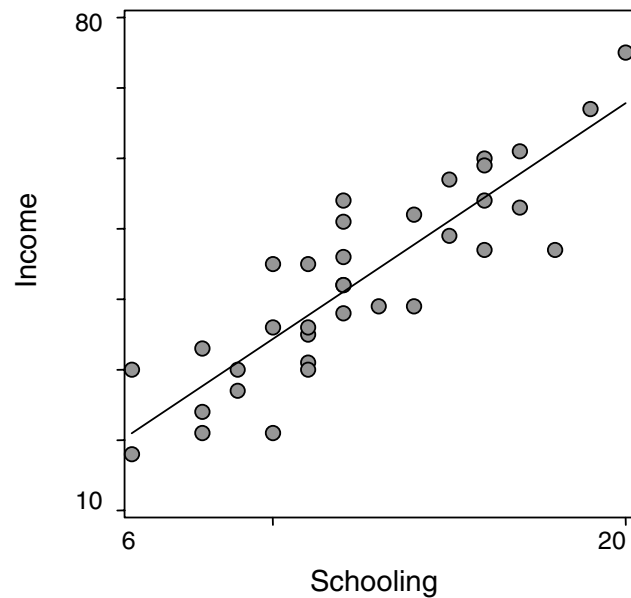
Social scientists often want to examine the relationship between two variables. To illustrate, an empirical example is unfolded. Imagine that educational sociologists in a small midwestern college seek information on the annual earnings of their students well after graduation. They want to know many things, including the influence of basic demographic forces, such as the socioeconomic background of the students' families. In particular, they seek to establish the link, if any, between the educational attainment of parents and the later job earnings of the child. To gather the necessary data on these and other research questions, they draw a RANDOM SAMPLE from the alumni list of 350 students who graduated 15 years ago. Because of cost considerations, they sample 1 out of 10, yielding a sample size of 35. Table 1 contains some of the data from the survey administered to them. In column 1 is the CODE number given the student. In column 2 is the parent education variable, measured as the number of years of formal schooling completed by the most educated parent (mother or father). In column 3 is the student income variable, measured as the reported gross annual income (in thousands of dollars) the student earned 15 years after graduation. In column 4 is the gender of the student, scored 1 = male or 0 = female. These data are listed as they would appear when entered into a typical computer STATISTICAL PACKAGE.

The research question at hand is whether parent education (variable X_1) helps account for later student income (variable Y). The dominant HYPOTHESIS is that they are positively related. The more educated the parental environment, the more child opportunities and incentives for learning and advancement and, ultimately, higher income on the job. Does an analysis of the data support the hypothesis? In bivariate regression, the first step is inspection of a SCATTERPLOT, as in Figure 1.

Table 1 A Data Set of Three Variables

	Variable 1	Variable 2	Variable 3
Case	Schooling	Income	Gender
1	6	18	0
2	6	30	0
3	8	21	0
4	8	24	0
5	8	33	1
6	9	27	0
7	9	30	0
8	10	36	1
9	10	45	1
10	10	21	0
11	11	35	0
12	11	45	1
13	11	31	0
14	11	30	1
15	11	36	0
16	12	42	0
17	12	51	1
18	12	54	1
19	12	38	1
20	12	46	0
21	12	42	1
22	13	39	0
23	14	52	1
24	14	39	0
25	15	57	1
26	15	49	1
27	16	60	1
28	16	47	0
29	16	59	1
30	16	54	0
31	17	61	1
32	17	53	1
33	18	47	0
34	19	67	1
35	20	75	0

Each dot, or point, in the plot represents the scores of a particular student on variables X_1 and Y . For example, student number 35 has a parent education score of 20 years and an income of \$75,000. She is represented by the dot in the upper right-hand corner. Indeed, the point for each student locates itself at the intersection of an imaginary perpendicular line from his or her Y value (on the vertical axis) and of an imaginary perpendicular line from his or her X value (on the horizontal axis). Visual inspection of the scatter of points suggests that X_1 is positively related to Y ; that is, lower values of X_1 tend to appear in the lower left-hand corner with lower values of Y , higher values of X_1 tend to appear in the upper right-hand corner with higher values

**Figure 1** Scatterplot With Regression Line

of Y . Overall, the constellation of dots, to the extent it suggests any geometric form, suggests a linear one. Put another way, a straight line would seem able to touch more of the points than any plausible curve. The formula for a straight line is $Y = a + bX$. The letter a is the CONSTANT value, where the line intercepts the y -axis, and the letter b denotes the SLOPE of the line. In Figure 1, a straight line is actually fitted to this scatterplot and expressed in the regression equation, predicted $Y = 0.88 + 3.35X_1$.

The regression line in Figure 1 is not arbitrarily drawn. Rather, it is the line of best fit, calculated from the LEAST SQUARES principle. Observe that some points fall on the line, and others do not. One sees positive errors (above the line) and negative errors (below the line). If all these errors (measured as the vertical distance of each point to the line) were simply added up, they would sum to zero, with the signs (+ and -) canceling out. However, when these errors are squared, there is no canceling out, and a measure of total error is produced, the sum of the squares of these errors (SSE). The sum of these squared errors is "least," the lowest possible value, hence the phrase "least squares." From calculus, the unique values for the intercept (a) and slope (b) that minimize the SSE are derived, and they constitute the ordinary least squares (OLS) solution. In our example, no other combination of INTERCEPT and slope values would produce a line that fit the data so well as the combination (0.88, 3.35) does.

The regression line serves as a PREDICTION EQUATION. Say the symbol for a predicted Y value is $\hat{Y}$, and $X = 12$. Then,

$$\begin{aligned}\hat{Y} &= 0.88 + 3.35X \\ &= 0.88 + 3.35(12) \\ &= 0.88 + 40.2 \\ \hat{Y} &= 41.08.\end{aligned}$$

The model predicts the student should earn about \$41,000 annually, if the highest educational attainment of the parent was 12 years of schooling. Suppose now that $X = 13$.

$$\begin{aligned}\hat{Y} &= 0.88 + 3.35(13) \\ &= 0.88 + 43.55 \\ \hat{Y} &= 44.43.\end{aligned}$$

One observes that when the educational attainment variable went up one year, the predicted value of income went up by 3.35, the slope value. This illustrates the general interpretation of the slope: b indicates the expected change in Y for a unit change in X_1 . The intercept also has a general interpretation: a indicates the EXPECTED VALUE of Y when $X_1 = 0$. When there are no X_1 values at 0, as is the case with these data, the intercept has a mathematical role (it is a constant that must be added to complete a prediction) but no substantive meaning.

The regression line can be used for predictions, but they are not perfect, as the points distant from the line demonstrate. How well the linear model actually fits the data is measured with the R -SQUARED (R^2), which assesses the variation in the dependent variable accounted for by the independent variable. It is a summary statistic ranging from 1.0 to .00. In the example, $R^2 = 0.76$, suggesting parent educational background accounts for 76% of the variation in student income.

MULTIPLE REGRESSION

Analysis with multiple regression extends the foregoing, by allowing for multiple independent variables. This is an important extension, for two reasons. First, a more complete explanation of a phenomenon becomes possible, because multiple causes, predictors, or determinants can be entertained simultaneously. Second, the specific effect of a particular X on Y can be better

understood, because the influence of other, perhaps CONFOUNDING, variables can be removed through statistical controlling. Consider the example. Other factors besides parent education undoubtedly help shape student income. A multiple regression model is in order, to take these other variables into account. There is a measure on at least one of these other variables—gender. Therefore, the revised theory is that student income (Y) is influenced by parent education (X_1) and gender (X_2). Here are the OLS estimates for this multiple regression model:

$$\hat{Y} = 0.44 + 3.12X_1 + 6.9X_2.$$

On the basis of these results, the interpretation is somewhat altered. The impact of an additional year of parent education appears slightly diminished, compared with the earlier bivariate result ($3.12 < 3.35$). This reduction in the slope coefficient can be attributed to statistically “holding constant” X_2 . Furthermore, gender itself seems to have an independent effect on income. Specifically, males can expect to earn about \$7,000 more than females, once parent education differences are controlled for. Overall, the $R^2 = 0.82$, indicating that these two independent variables together account for 82% of the variation in income. This is a boost in the R^2 of six percentage points, moving from the bivariate to the multivariate model (a gain that drops a bit with the ADJUSTED R -SQUARED).

ASSUMPTIONS

Regression analysis of nonexperimental social science data aims to reveal structural links—causal, explanatory, predictive—between variables in the real world. More formally, the goal is to estimate POPULATION parameters, the fixed intercept and slope values that reliably join the variables under study. These estimates are accurate to the extent that key assumptions are met. If the assumptions are ignored, regression results may have no reality beyond the numbers typed on a page. Granting the Table 1 data, do the regression estimates presented reveal the actual connection between student income, parent education, and gender? Yes, to the extent that they are supported by the assumptions presented below.

Most regression analysis is carried out on a SAMPLE, rather than on the target population as a whole. With estimation of the sample equation, say, $Y = a + b_1X_1 + b_2X_2 + e$, inferences are made to the

parameters in the population equation, say, $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. The sample must be a scientific PROBABILITY SAMPLING of sufficient size. Furthermore, because the samples are only samples, SIGNIFICANCE TESTING must be applied, to rule out chance results. For example, with the sample of 35 students out of the population of 350, we observe a slope coefficient of 3.12 in the multiple regression equation. A significance test (.05, two-tailed) applied to that regression coefficient allows us to reject the NULL HYPOTHESIS that the link between parent education and income is zero in the population, that is, $\beta_1 = 0$.

Besides assumptions of proper sampling and significance testing, there are the classical linear multiple regression assumptions, which can be variously stated. Broadly, these include no SPECIFICATION error, no measurement error, no perfect COLLINEARITY, and no error term problems. With respect to specification, the essential ideas are that the model has the right independent variables, and their relationship to the dependent variable is linear. With respect to measurement, the variables should be quantitative and accurately scored. With respect to collinearity, no independent variable is allowed to be a perfect linear function of all the others. With respect to the error term, it should not be related to the independent variables either as variance (to preserve HOMOSKEDASTICITY) or as a variable (because the data are nonexperimental), and it should not be related to itself (to avoid AUTOCORRELATION). When these assumptions are met, the OLS estimator is BLUE, or BEST LINEAR UNBIASED ESTIMATOR. Certain analysts prefer to think of the classical linear regression assumptions as lying along a continuum, say, from 1 to 10, where 1 means *not met at all* and 10 means *perfectly met*. In that view, the closer the results to 10, the closer the inferences are to reality.

OLD DIRECTIONS AND NEW

Sir Francis Galton (1822–1911) can perhaps be considered the founder of regression analysis, because he was the first to perceive that a straight line could make sense of the whirl of points in a scatterplot. He and his disciple Karl Pearson (1857–1936) made extensive investigations of scatterplots relating the characteristics of fathers and sons. For example, Pearson examined the heights of 1,078 father-and-son combinations and found that very tall fathers tended to have shorter sons. This was “to regress toward the mean,” or what Galton called “regression to mediocrity.” The straight

line that depicted this regression effect was named the “regression line.” Modern regression and CORRELATION methods were launched by the papers published on genetics in the journal *Biometrika*, particularly the work of Pearson coming out in 1903.

OLS has been the central method of regression estimation. The least squares method was discovered independently by French mathematician Adrien Marie Legendre (1752–1833) and German mathematician Carl Friedrich Gauss (1777–1855). When the classical multiple regression assumptions are met, OLS estimators cannot be improved upon. A great deal of work, especially since the 1960s, has gone into REGRESSION DIAGNOSTICS, which attempt to test assumptions and provide corrections. However, certain models are intrinsically nonlinear, and their parameters cannot be efficiently estimated with OLS. A case in point is when the dependent variable is dichotomous, say, a “yes” versus “no” vote choice. Here, a MAXIMUM LIKELIHOOD ESTIMATION technique, such as LOGISTIC REGRESSION, is generally preferred. Logistic regression approaches, including POLYTOMOUS ones, have recently been a vigorous area of research. Although these approaches are, in some ways, very far from OLS, the logic of the modeling, and many of the problems of inference due to violation of assumptions, is the same.

—Michael S. Lewis-Beck

REFERENCES

- Kmenta, J. (1997). *Elements of econometrics* (2nd ed.). Ann Arbor: University of Michigan Press.
- Lewis-Beck, M. S. (1980). *Applied regression: An introduction*. Beverly Hills, CA: Sage.
- Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. W. (1996). *Applied linear regression models*. New York: Irwin.

REGRESSION COEFFICIENT

The term *REGRESSION* refers to a group of popular methodologies that aims to explain variation in a DEPENDENT VARIABLE (Y) using a set of k INDEPENDENT VARIABLES (X_k). In the context of ORDINARY LEAST SQUARES (OLS), a regression coefficient ($\hat{\beta}_k$) is associated with each independent variable ($X_1, X_2, \dots, X_k$) and gives the expected direct effect of a one-unit change in that independent variable on the dependent variable, holding all other independent variables constant. The regression coefficient is

always expressed in units of the dependent variable. In addition, an INTERCEPT term ($\hat{\beta}_0$) is usually estimated and yields the expected value of the dependent variable if all independent variables were equal to zero.

In OLS estimation with two independent variables (X_1 and X_2) and n observations, the formulas to calculate the intercept and regression coefficients are as follows:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2,$$

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^n (y_i - \bar{y})(x_{1i} - \bar{x}_1) \right) \left(\sum_{i=1}^n (x_{2i} - \bar{x}_2)^2 \right) - \left(\sum_{i=1}^n (y_i - \bar{y})(x_{2i} - \bar{x}_2) \right) \left(\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)}{\left(\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 \right) \left(\sum_{i=1}^n (x_{2i} - \bar{x}_2)^2 \right) - \left(\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2}.$$

Consider the following hypothetical example:

Table 1 Calculation of Regression Coefficients

<i>I</i>	<i>y</i>	<i>x</i> ₁	<i>x</i> ₂	
1	22	39	6	$\bar{y} = 24.6$
2	21	41	22	$\bar{x}_1 = 35.2$
3	24	39	20	$\bar{x}_2 = 14.6$
4	24	34	8	$\sum_{i=1}^n (y_i - \bar{y})(x_{1i} - \bar{x}_1) = -82.1$
5	26	33	21	$\sum_{i=1}^n (y_i - \bar{y})(x_{2i} - \bar{x}_2) = 21.2$
6	32	27	20	$\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = -31.1$
7	20	31	16	$\sum_{i=1}^n [(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)]^2 = 967.9$
8	27	32	8	$\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 = 197.6$
9	25	41	10	$\sum_{i=1}^n (x_{2i} - \bar{x}_2)^2 = 338.2$

Based on these data, and the formulae, we get the following model estimation:

$$\hat{y} = 38.7 - 0.412x_1 + 0.025x_2.$$

The intercept term indicates that if x_1 and x_2 were both zero, we would expect Y to equal 38.7. This may, or may not, have substantive meaning in practical research. The coefficient for x_1 denotes that for every one-unit change in x_1 , we would expect a 0.412-unit decrease in the dependent variable, holding x_2 constant. Similarly, the coefficient for x_2 shows that for every one-unit change in that variable, we would expect Y to increase by 0.025 units, holding x_1 constant. If none of the ASSUMPTIONS of the OLS model is violated, the estimated coefficients are UNBIASED and maximally efficient estimates of the true POPULATION parameters. Additionally, all regression coefficients have associated STANDARD ERRORS that enable us to

test the hypothesis that they are statistically different from zero.

The example illustrates the intuition behind regression coefficients and shows how to calculate and interpret them in a simple model. The calculation procedures and interpretation of regression coefficients are somewhat different in more complicated models, such as models with INTERACTIONS and NONLINEAR relationships between the independent variables, and models with limited dependent variables. In limited dependent variable cases, the regression coefficients are often calculated via MAXIMUM LIKELIHOOD ESTIMATION and cannot be interpreted as having a linear effect on the dependent variable (Greene, 2003). The basic intuition, however, is the same—a regression coefficient tells us how one variable affects another while holding other explanations constant.

—Kevin J. Sweeney

REFERENCES

- Achen, C. H. (1982). *Interpreting and using regression*. Beverly Hills, CA: Sage.
- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.

REGRESSION DIAGNOSTICS

Common violations of the assumptions of LEAST SQUARES regression are nonlinearity, HETEROSKEDASTICITY, and non-normally distributed error terms. Unusual observations can also have undue influence on the regression equation, and MULTICOLLINEARITY can cause unstable estimates. It is important to perform diagnostics of the model because if these problems are not rectified, OLS estimates are no longer the BEST LINEAR UNBIASED ESTIMATORS.

Nonlinearity in simple regression is easily assessed by examining a SCATTERPLOT of the dependent variable Y against the independent variable X . In multiple regression, however, simple scatterplots no longer suffice because they show only the marginal relationships between Y and each of the individual X s, when the model gives PARTIAL REGRESSION COEFFICIENTS. Partial residual plots (also called component-plus-residual plots) are thus used to assess linearity in the

multivariate case. The partial residual $E(j)$, for the j th independent variable, is simply the linear component of the partial regression of Y on X_j added to the least squares residuals. The $E(j)$ are plotted versus X_j , and linearity is assessed in the same manner as the scatterplot for simple regression.

Nonlinear relationships can be handled by transforming Y and/or X (when the nonlinearity is simple and monotone), POLYNOMIAL EQUATIONS (when the relationship is nonmonotone but has relatively few changes in direction), or LOCAL REGRESSION (when the nonlinear relationship is very complex).

Unusual observations that are OUTLIERS (i.e., have an unusual Y value given their X value) and have high leverage (characterized by an unusual X value) unduly influence the coefficients of a regression model.

Statistical tests for outliers are usually based on the studentized residuals (sometimes called standardized residuals). The “mean shift” regression model—which simply includes a DUMMY VARIABLE coded 1 for the unusual observation and 0 for all other cases—is a commonly used method for finding the studentized residual for an outlier. If the coefficient for the dummy variable is statistically significant, the observation significantly deviates from the pattern of the rest of the data. Nevertheless, because we select the furthest outlier, rather than observations at random, a simple t -test is inappropriate. A Bonferroni adjustment to the p value remedies this problem. The Bonferroni p value for the largest outlier is $p = 2np'$, where p' is the unadjusted p value from a t -test with $n - k - 2$ degrees of freedom.

Leverage is measured by the hat value h_i , which captures how much the observed Y_i affects the predicted value $\hat{Y}$. In simple regression, hat values are determined by $h_i = \sum_{j=1} (X_j - \bar{X})^2$ and are bounded between $1/n$ and 1, with an average of $(k + 1)/n$. Hat values that are approximately twice the average are usually considered noteworthy.

The most commonly used diagnostic for INFLUENTIAL CASES is Cook’s Distance (or, more simply, Cook’s D) which measures the impact of an unusual observation on the slope coefficient. Cook’s D is simply calculated:

$$D_i = \frac{E_j^{*2}}{k + 1} \times \frac{h_i}{1 - h_i}.$$

There is no universally accepted cutoff for Cook’s D, but a loose rule of thumb is that values greater than $4/(n - k - 1)$ should be looked at closely.

When trying to identify unusual cases, it is helpful to examine the data graphically. In the simple regression case, scatterplots are useful. Added-variable plots (also called partial regression plots) are useful methods of assessing joint influence in multiple regression. In all cases, influence plots (also called “bubble plots”)—which plot hat values, studentized residuals, and Cook’s Distances on a single graph—can be particularly useful.

Unusual observations may reflect miscoding, in which case the observations can be rectified or deleted. Sometimes, however, an outlier is of substantive interest, so we may decide to deal with it separately from the rest of the analysis. The presence of many outliers may indicate that an important explanatory variable is missing from the model. If there are no strong reasons to remove outliers, ROBUST regression, which downweights influential data, can be used in place of OLS.

HETEROSKEDASTICITY is usually assessed using a residual plot, which plots the residuals E_i (or studentized residuals, E_i^*) against $\hat{Y}$ (the predicted value of Y). If the values of Y are all positive, we can use a spread-level plot, which plots the $\log|E_i^*|$ (called the log spread) against the $\log \hat{Y}$ (called the log level). The SLOPE, B , of the regression line fit to the spread-level plot suggests the variance-stabilizing power transformation p , for Y , with $p = 1 - B$. If the error variances are proportional to a particular X , WEIGHTED LEAST SQUARES provides a good alternative to OLS.

Non-normality can be diagnosed using quantile comparison plots, which plot the distribution of the studentized residuals E_i^* against the quantiles of the unit-normal distribution or the t -distribution. If the distribution is not skewed, the relationship will be roughly linear. Modality is best assessed using HISTOGRAMS or density estimates of the residuals. Transformations of either Y or the X s can often remedy a heavy-tailed error distribution, whereas model respecification—that is, including a missing discrete predictor—can sometimes fix a multimodal problem.

Multicollinearity can be indicated by a very high pairwise CORRELATION (.8 or higher) between two independent variables, but it is also possible in the absence of high pairwise correlations because it reflects the multiple correlations between each X and all the other X s in the model. It is more effective, then, to

examine the R^2 s for each X regressed on all other X s in the model ($R^2 = .6$ or higher is of concern). Related to the R^2 , the VARIANCE-INFLATION FACTORS (VIF) for each variable, which is defined by $VIF = 1/(1 - R_j^2)$, indicates the impact of collinearity on the precision of the SLOPE, B_j . The square root of the VIF is the factor by which the standard error is inflated because of the multicollinearity.

—Robert Andersen

REFERENCES

- Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart.
- Cook, D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425), 177–189.
- Cook, D. R., & Weisberg, S. (1999). *Applied regression including computing and graphics*. New York: Wiley.
- Fox, J. (1991). *Regression diagnostics* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–079). Newbury Park, CA: Sage.
- Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
- Fox, J. (2002). *An R- and S-PLUS companion to applied regression*. Thousand Oaks, CA: Sage.

REGRESSION MODELS FOR ORDINAL DATA

Ordinal-level variables (such as LIKERT SCALES from survey questions, some measures of social class, etc.) are commonly used as DEPENDENT VARIABLES in the social sciences. Such data can be modeled in several ways within a REGRESSION framework: (a) The ordering of the categories can be ignored, allowing the use of binary or MULTINOMIAL LOGIT models; (b) the categories can be treated as though they were measured on an interval scale, enabling the use of ORDINARY LEAST SQUARES (OLS); and (c) models developed specifically for ordinal dependent variables can be used. Each of these methods has strengths and limitations.

The first strategy—ignoring the ordered nature of the dependent variable—is most feasible when there are few categories to the dependent variable. For example, if the categories of the dependent variable can be legitimately collapsed into two categories without losing too much information, simple binary logit models (LOGISTIC REGRESSION) can be used. For these models, ODDS RATIOS can be used to determine

differences between categories. The limitation of this approach, however, is that by collapsing categories, we lose possibly important information about the dependent variable. Although multinomial logit models can be used if we choose not to collapse the categories, these models are inefficient and become increasingly difficult to interpret as the number of categories in the dependent variable rises because of the large number of parameters that are estimated.

The strength of the second strategy—using OLS—is that the parameter estimates are easily interpreted (see SLOPE). OLS can be legitimately used if the distribution of the dependent variable appears to be roughly normal and the ordered categories accurately represent an underlying continuous distribution. If the variable has many categories, a HISTOGRAM can give a rough assessment of the distribution. On the other hand, if the variable is measured using only a few categories, the variable will be characterized by such a high degree of measurement error that this assumption cannot be reasonably assessed. These models can also be problematic if prediction is a main goal of the research, because predicted values outside the range of the variable are possible. Nonetheless, because the variable is knowingly measured with error, having predicted values outside the range of the variable is not necessarily problematic, especially if the research is primarily concerned with determining relationships rather than making predictions.

Because we often cannot assume that the distances between categories of ordered variables are interval—meaning that they do not satisfy the criteria of OLS (in particular, the assumption of constant error variance)—regression models specifically for ordinal data have been developed. These models include the ordered logit model, which is a generalization of logistic regression, and the ordered PROBIT ANALYSIS model, which is a generalization of the binomial probit model. Here, we will deal with the most commonly used of the ordered logit models: the proportional-odds model.

Logits can directly incorporate the ordering of categories in a dependent variable, resulting in a model with a simpler interpretation and potentially greater power than multinomial logit models, which are typically used for unordered categorical dependent variables. Consider a latent continuous variable ξ , which is a linear function of the explanatory variables, the X s, plus a random error:

$$\xi_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i.$$

Assume here that ξ represents attitudes toward abortion and the β s are the regression slopes. Assume also that we have only a single 5-point LIKERT SCALE to tap attitudes toward abortion, Y . In this case, the latent continuous variable ξ has been divided into five categories, $m = 5$, with $m - 1 = 4$ boundaries, $\alpha_1 < \alpha_2 < \dots < \alpha_{m-1}$:

$$Y_i = \begin{cases} 1 & \text{if } \xi_i \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < \xi_i < \alpha_2 \\ \vdots & \\ m-1 & \text{if } \alpha_{m-2} < \xi_i \leq \alpha_{m-1} \\ m & \text{if } \alpha_{m-1} < \xi_i. \end{cases}$$

The cumulative probability distribution of Y is then easily found:

$$\begin{aligned} \Pr(Y_i \leq j) &= \Pr(\xi_i \leq \alpha_j) \\ &= \Pr(\alpha + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i \leq \alpha_j) \\ &= \Pr(\varepsilon_i \leq \alpha_j - \alpha - \beta_1 X_{i1} - \dots - \beta_k X_{ik}). \end{aligned}$$

From these, we can find the ordered LOGIT MODEL:

$$\begin{aligned} \text{logit}[\Pr(Y_i > j)] &= \log_e \frac{\Pr(Y_i > j)}{\Pr(Y_i \leq j)} \\ &= (\alpha - \alpha_j) + \beta_1 X_{i1} + \dots + \beta_k X_{ik}. \end{aligned}$$

Here, the logits are cumulative logits that contrast categories above category j with category j and below it. Only one slope coefficient is found for each X , but there is a separate intercept for each category. The first category (or last, depending on the software) is set as a reference category to which all the intercepts relate.

Coefficients from ordered logit models are interpreted in the same way as for the dichotomous logit model, except rather than referring to a single baseline category, we contrast each category and those below it with all the categories above it. In other words, the logit tells us the log of the odds that Y is in one category and below versus above that category.

Despite the appealing parsimony, the proportional-odds models must be used cautiously. These models assume that the regression slopes are equal (referred to as the proportional-odds assumption or the assumption of parallel slopes), something that does not often hold in practice, especially with large samples. A simple way to test whether this assumption is met is to compare the fit of the model with the fit of the multinomial logit model, which can be seen as a more general case. The difference in deviance of the two models

gives a generalized LIKELIHOOD RATIO test, which is distributed as chi-square with DEGREES OF FREEDOM equal to the difference in the number of parameters between the multinomial logit model and the proportional-odds logit model.

—Robert Andersen

REFERENCES

- Agresti, A. (1984). *An introduction to categorical data analysis*. New York: Wiley.
- Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
- McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42, 109–142.
- Menard, S. (2002). *Applied logistic regression analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–106). Thousand Oaks, CA: Sage.

REGRESSION ON ...

The phrase “regression on” refers to the act of projecting the INDEPENDENT VARIABLE vector (sometimes referred to as the “regressand”) onto the Euclidean subspace spanned by the DEPENDENT VARIABLES (sometimes collectively referred to as “REGRESSORS”). Although the phrase refers to the underlying geometry of least squares, it is often used to indicate the set of independent variables used to model the observed variation in the dependent variable.

For example, if a researcher is interested in the simple LINEAR REGRESSION of $Y = b_0 + b_1 X + e_1$, we say that “ Y is regressed on X ”—indicating that the simple regression recovers the set of coefficients $\{b_0, b_1\}$ such that the projection of Y onto the subspace spanned by X (in this case, a line) is given by $b_0 + b_1 X$. Because different values of b_0 and b_1 (say, b'_0 and b'_1) project Y onto the subspace of X differently (i.e., $b'_0 + b'_1 X$), there are (infinitely) many possible projections/regressions. To determine which projection/regression is of most interest, researchers must assess the “fit” according to a given criterion. A commonly used criterion is that of LEAST SQUARES, which denotes the “best” projection as the choice of $\{b_0, b_1\}$ for which $|Y - (b_0 + b_1 X)|^2$ is minimized.

In the MULTIPLE REGRESSION ANALYSIS context, “regression on” refers to the projection of the dependent variable vector into the subspace spanned by the set of independent variables—with the REGRESSION

PLANE defining the plane that minimizes the squared differences between the observed values of the dependent variable and its projection into the regression plane.

—Joshua D. Clinton

REFERENCE

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.

REGRESSION PLANE

The regression (hyper) plane represents the projection of the dependent variable into the Euclidean subspace spanned by the independent variables.

MULTIPLE REGRESSION using the LEAST SQUARES criterion involves finding the hyperplane that minimizes the sum of squared differences between the observed values of the dependent variable, and the projection of the dependent variable onto that hyperplane.

The hyperplane that minimizes this difference is called the regression hyperplane.

An example clarifies the concept. Suppose that a researcher is interested in the (linear) relationship between the dependent variable Z and the independent variables X and Y . The equation of interest is therefore

$$Z = b_0 + b_1X + b_2Y + e.$$

Consider the nature of the data being fitted—each observation i defines a point in the three-dimensional space (X_i, Y_i, Z_i) . Intuitively, in the above example, the regression plane is the two-dimensional flat surface that “best fits” the set of three-dimensional data points—where “best fits” means that the regression plane minimizes the sum of squared differences (i.e., residuals) between the surface of the plane (which represents the predicted value of Z given X and Y) and the observed Y . For this example, the regression plane defines the two-dimensional plane that minimizes the sum of squared residuals between the independent variable Z and the plane. More generally, if there are k dependent variables (excluding the constant term), the regression hyperplane is the k -dimensional flat surface

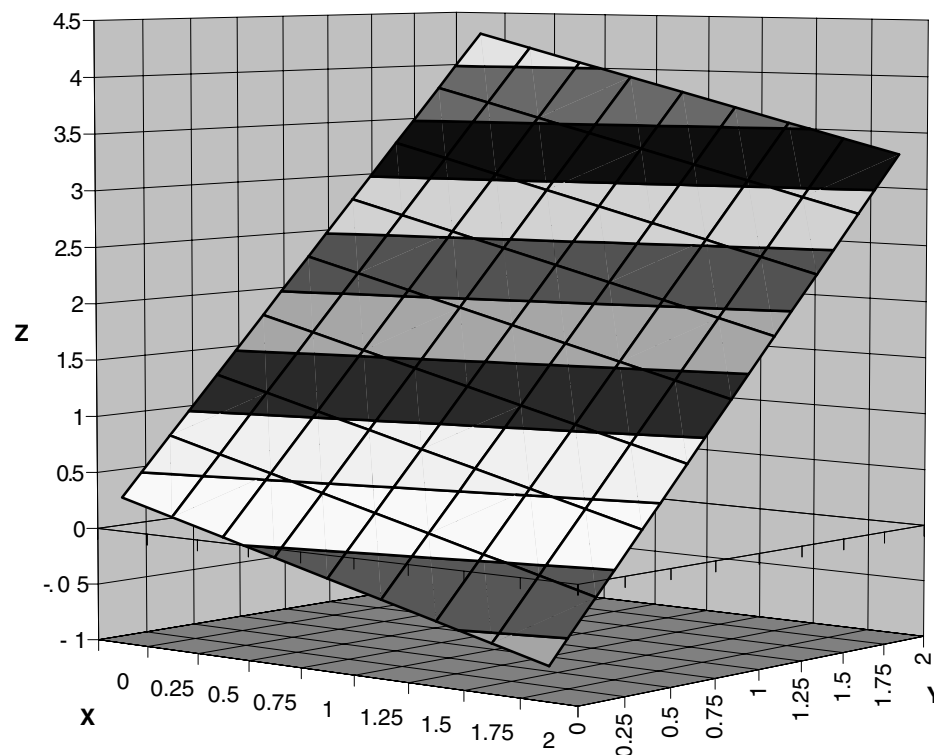


Figure 1 Regression Plane

representing the best linear projection of Y into the $(k + 1)$ -dimensional space spanned by the k independent variables plus the dependent variable. The regression plane is a special case for $k = 2$.

REGRESSION COEFFICIENTS define the slope of the regression plane in the dimension defined by the respective independent variable. For example, in the three-space defined above, b_0 defines the “height” of the regression plane in the third (i.e., Z) dimension when $X = Y = 0$, b_1 defines the slope of the regression plane in the first dimension (i.e., the angle at which the plane intersects the Z - X plane), and b_2 defines the slope of the regression plane in the second dimension (i.e., the angle at which the plane intersects the Z - Y plane). The equation for the regression plane is $b_0 + b_1X + b_2Y$. In other words, just as simple regression involves finding the best fitting line, the generalization to multiple regression involves finding the best fitting plane.

An illustration clarifies. The plane depicted in Figure 1 is characterized by the regression equation $Z = 0.3 - 0.5X + 2Y$. Because the regression plane describes the mapping of Z onto X and Y , it is possible to generate a predicted value of Z for any (X, Y) pair. The regression plane depicts these predictions. The geometric interpretation of the slope coefficients is also evident from the figure. The regression coefficient for X of -0.5 implies that the plane in the figure “tilts” to the left; higher values of X lead to lower predicted values of Z . In contrast, the plane “rises” in the back because the slope coefficient for Y is positive. The intercept (in this case, 0.3) is given by the “height” of the plane when $X = Y = 0$.

—Joshua D. Clinton

REFERENCE

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.

REGRESSION SUM OF SQUARES

In LINEAR REGRESSION analysis, it is useful to report how well the sample regression line fits the data. A typical measure of GOODNESS OF FIT for a linear regression model is the fraction of the sample variation

in the response (or dependent) variable (Y) that is explained by the variations in independent variables (X s). Regression sum of squares (RSS) is the key component of a common goodness-of-fit measure for a linear regression model. RSS tells data analysts how much of the observed variation in Y can be attributed to the variation in the X s. In other words, RSS indicates the jointly estimated effect of X s on the variation of Y .

To explain the idea of RSS, it is useful to define SAMPLE variation in the DEPENDENT VARIABLE Y . The observed variation in Y for a specific observation, say, Y_i , is simply its deviation from the mean of Y or, symbolically, $Y_i - \bar{Y}$. Because we are interested in all observations, a convenient summary measure is the sum of squared deviation for all observations. This gives total variation of Y or total sum of squares (TSS):

$$\text{TSS}(\text{Total Variation of } Y) = \sum_i (Y_i - \bar{Y})^2.$$

The question that arises now is how to decompose the total variation in Y , which is also known as “total sum of squares” and denoted by TSS, so that we can measure how much of the total variation in Y can be attributed to the variation in the X s. We can consider the following identity, which holds for all observations:

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}), \quad (1)$$

where $\hat{Y}_i$ is the fitted value of Y for a specific observation with an independent variable that equals X_i . That is, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ for a SIMPLE REGRESSION model.

We already defined the left-hand side of the equal sign in equation (1) as the deviation of a specific value of Y (i.e., Y_i) from the mean of Y . The first component on the right-hand side gives the RESIDUAL e_i because, by definition, $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i = \hat{Y}_i + e_i$. Thus, $Y_i - \hat{Y}_i = e_i$. The second right-hand term gives the difference between the fitted (or predicted) value of Y and the mean of Y . Because $\hat{Y}_i$ is predicted by using the values of the independent variable (X_i), the second right-hand term represents the part of the variation in Y that is explained by the variation in X . The decomposition of variation in Y for a simple regression (i.e., only one independent variable) is shown in Figure 1.

As noted previously, equation (1) applies for only a single observation. We can obtain a summary measure for all of the data by squaring both sides of the equality

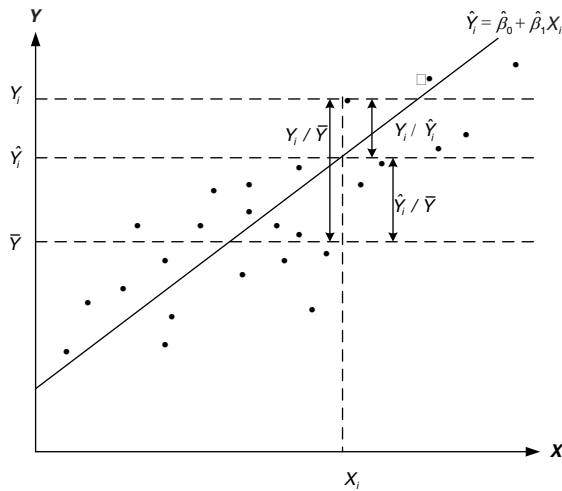


Figure 1 Decomposition of Variation in Y

and summing over all sample observations. This gives the following equation:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}). \quad (2)$$

The last term in equation (2) is identically zero because of two ASSUMPTIONS of a classical normal linear REGRESSION model, zero mean of the disturbance [$E(e_i|X_i) = 0$] and zero covariance between disturbance and independent variables [$\text{cov}(e_i, X_i) = E(e_i X_i) = 0$]. As a result, equation (2) can be rewritten as

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2.$$

Total Sum of Squares (TSS)	=	Error Sum of Squares (ESS)	+	Regression Sum of Squares (RSS)
----------------------------	---	----------------------------	---	---------------------------------

(3)

Equation (3) shows that regression sum of squares (RSS) can be defined as the sum of squared difference between the predicted value of the dependent variable ($\hat{Y}_i$) and the mean of the dependent variable ($\bar{Y}$) for all observations $i = 1, 2, \dots, n$. Equation (3) also shows that the variation in Y is partly due to the effect of random disturbance (i.e., the ESS term) and partly due to the changes in X (i.e., the RSS term).

The proportion of total variation in Y (or total sum of squares, TSS) that can be attributed to the variation in X is a commonly used measure of the goodness of fit, which is known as the (sample) COEFFICIENT OF DETERMINATION and denoted by R -SQUARED (R^2).

Note that R^2 is a nonnegative quantity and cannot be greater than 1 (i.e., $0 \leq R^2 \leq 1$). More information on other properties of R^2 and technical derivation regarding how to obtain equation (3) may be found in Fox (1997); Rawlings, Pantula, and Dickey (1999); Kmenta (1986); and Johnston and Dinardo (1997).

—Cheng-Lung Wang

REFERENCES

- Fox, J. (1997). *Applied regression analysis: Linear models and related methods*. Thousand Oaks, CA: Sage.
- Johnston, J., & Dinardo, J. (1997). *Econometric methods* (4th ed.). New York: McGraw-Hill.
- Kmenta, J. (1986). *Elements of econometrics* (2nd ed.). New York: Macmillan.
- Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1999) *Applied regression analysis: A research tool*. New York: Springer.

REGRESSION TOWARD THE MEAN

First noted by Francis Galton (1822–1911), regression toward the mean (RTM) is a statistical phenomenon that occurs between any two imperfectly correlated variables. The variables could be either distinct but measured contemporaneously (e.g., height and weight), or the same and measured at different times (t and $t + 1$). In either case, extreme observations on the first variable are likely to be less extreme—that is, closer to the mean—on the second variable. Galton's work focused on the hereditary size of pea seeds and human beings. He found that both “regressed toward mediocrity”; larger parents had slightly smaller offspring, whereas the smaller parents had slightly larger offspring. When RTM is ignored, it is a threat to the INTERNAL VALIDITY of experiments.

RTM occurs whenever two variables are not perfectly correlated ($\rho < 1$). If two variables are STANDARDIZED (both have a mean = 0 and a standard deviation = 1) and NORMALLY DISTRIBUTED, we can use their correlation to estimate the expected RTM:

$$E(Z_Y|Z_X) = \rho Z_X.$$

For example, if the weights of mothers and daughters are standardized, and their correlation is .8, we would expect a mother who was 2 STANDARD DEVIATIONS above the “mother” mean to have a daughter who was 1.6 standard deviations above the “daughter” mean. Importantly, the formula indicates that the expected amount of RTM increases as the correlation between the variables decreases. RTM is also bidirectional:

$$E(Z_X|Z_Y) = \rho Z_Y.$$

So, we would expect a daughter who was 2 standard deviations above the “daughter” mean to have had a mother who was only 1.6 standard deviations above the “mother” mean.

Although the formulas given above are useful for estimating the expected amount of RTM for an individual, RTM is inherently a group phenomenon. Any particular individual may move further from the mean upon second observation, but groups will always regress toward the mean as long as the correlation between the two variables is imperfect. Furthermore, the more extreme a group is, the more it will regress in absolute terms. Groups with means close to the sample mean do not have very far to regress, and the further the group mean is from the sample mean, the more room there is for regression. This has important implications for the internal validity of many social science experiments. Often, subjects are selected because they have performed particularly well or poorly. Especially in the latter case, an intervention may be performed to address the problem, and subjects will almost always show improvement. Given that the group is extreme, however, the improvement may be explicable by RTM alone.

—Kevin J. Sweeney

REFERENCES

- Campbell, D. T., & Kelly, D. A. (1999). *A primer on regression artifacts*. New York: Guilford.
- Galton, F. (1886). Regression toward mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, 246–263.
- Smith, G. (1997, Winter). Do statistics test scores regress toward the mean? *Chance*, pp. 42–45.

REGRESSOR

Regressor is a term that describes the INDEPENDENT VARIABLE(s) in a regression equation that explain

VARIATION in the DEPENDENT VARIABLE (regressand). The regressor is also known as the EXPLANATORY VARIABLE or the EXOGENOUS VARIABLE. In a system of equations, a variable that is the regressor in one equation may be the regressand in another, thus complicating the use of the term *exogenous*.

A regressor may be a QUANTITATIVE VARIABLE (i.e., measured on an INTERVAL or RATIO SCALE) or a QUALITATIVE VARIABLE (i.e., observed as a nominal or categorical trait). As a matter of convenience in the classical linear regression model, regressors are assumed to be nonstochastic or “fixed in repeated samples.” This assumption may be relaxed with little cost, particularly if the sample size is large.

—Brian M. Pollins

REFERENCES

- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Gujarati, D. N. (2003). *Basic econometrics* (4th ed.). Boston: McGraw-Hill.
- Koutsoyiannis, A. (1978). *Theory of econometrics: An introductory exposition of econometric methods* (2nd ed.). London: Macmillan.

RELATIONSHIP

A relationship is the connection between two or more CONCEPTS or VARIABLES. If, for example, two variables are related to each other, then they tend to move together in some nonrandom, systematic way. For example, as X scores go up, Y scores tend to go up. This link between X and Y can be quantified in a MEASURE OF ASSOCIATION, decided in part by the LEVEL OF MEASUREMENT of X and Y . The most prevalent measure of association is PEARSON’S CORRELATION COEFFICIENT, which assesses the linear relationship between X and Y , assuming these variables are measured at the INTERVAL level. A relationship can exist between three or more variables. For example, three variables— X_1 , X_2 , and X_3 —can be related to variable Y . If all the variables are interval, then a MULTIPLE CORRELATION coefficient, R , could be calculated to estimate this relationship.

—Michael S. Lewis-Beck

RELATIVE DISTRIBUTION METHOD

Differences among groups or changes in the distribution of a variable over time are a common focus of study in the social sciences. Traditional parametric models restrict such analyses to conditional MEANS and VARIANCES, leaving much of the distributional information untapped.

Relative distributional methods aim to move beyond means-based comparisons to conduct detailed analyses of distributional difference. As such, they present a general framework for comparative distributional analysis.

The relative DISTRIBUTION provides a graphical display that simplifies exploratory data analysis, a statistically valid basis for the development of hypothesis-driven summary measures, and location, shape, and covariate decompositions that identify the sources of distributional changes within and between groups.

The main idea of the relative distribution approach can be seen in an application to the comparison of earnings. The probability density functions (PDFs) in Figure 1(a) show the distribution of logged annual earnings for men and women who worked full-time, full-year in 1997. The women's distribution is clearly downshifted, but this graphical display provides little additional information useful for comparison.

The *relative distribution* is the set of percentile ranks that the observations from one distribution would have if they were placed in another distribution. In this example, it is the set of ranks that women earners would have if they were placed in the men's earnings distribution.

Figure 1(b) shows the density for the relative distribution of women's to men's earnings. The smooth line encodes the relative frequency of women to men at each level of the earnings scale; the value of this ratio is shown on the vertical axis. The top axis shows the dollar value of the log earnings, the bottom axis shows the rank in the men's distribution. The histogram encodes the fraction of women falling into each decile of the men's earnings distribution. We can see from the histogram that 20% of the women fall in the bottom decile of the men's earnings distribution, and another 18% in the second decile. In all, about 75% of women earn less than the median male (the sum of the first five deciles). By contrast, less than 5% of women's earnings reach the top decile of the men's distribution.

(a) Log Earnings PDFs for Men and Women, 1997



(b) Relative PDF, Women:Men

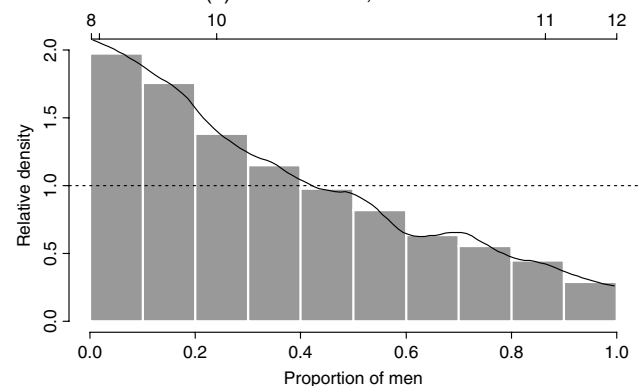


Figure 1

The differences between the men's and women's distributions can be divided into two basic components: differences in location and shape. If the women's earnings distribution is a simple downshifted version of the men's, then after matching the medians (or other location parameter), the two distributions should be the same. Differences that remain after a location adjustment are differences in distributional shape.

The relative distribution can be decomposed into these location and shape components, and Figure 2 shows the residual shape differences in men's and women's earnings. It is constructed by median-matching the women's earnings distribution to the men's (the median earnings ratio is about 1.4) and constructing the new relative distribution. After adjusting for median differences, the relative distribution is nearly flat, indicating that most of the difference between the two groups is due to the median downshift in women's earnings. There are some residual differences, however. Women's earnings have relatively less

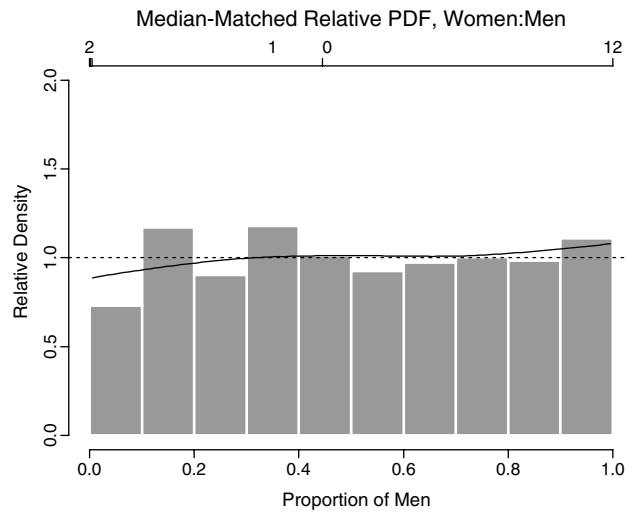


Figure 2

density in the lower tail (because the bottom decile is below 1) and relatively more density in the upper tail. This probably reflects the dramatic losses experienced by low-earnings men in the past 20 years, and the corresponding gains made by high-earnings women. The lower tail difference may also signal a minimum wage effect, as the women's median is lower to begin with, so their lowest earnings are closer to their own median than is the case for men. More information on these trends can be found in the review article of Morris and Western (1999). Handcock and Morris (1999) is a book-length treatment of relative distribution theory and includes historical references.

—Mark S. Handcock and Martina Morris

REFERENCES

- Handcock, M. S., & Morris, M. (1999). *Relative distribution methods in the social sciences*. New York: Springer-Verlag.
- Morris, M., & Western, B. (1999). Inequality in earnings at the close of the 20th century. *Annual Review of Sociology*, 25, 623–657.

RELATIVE FREQUENCY

One of the two primary reasons for applying statistical methods is to summarize and describe DATA. (The other reason is to make INFERENCES from the data.) *Relative frequency* is a simple tool to summarize

Table 1 Relative Frequency of Intelligence Quotient (Grouped Into Intervals of Five Units)

<i>Intelligence Quotient</i>	<i>Frequency</i>	<i>Relative Frequency</i>	<i>Percentage</i>
81–85	1	0.01	1.0
86–90	1	0.01	1.0
91–95	9	0.09	9.0
96–100	14	0.14	14.0
101–105	12	0.12	12.0
106–110	11	0.11	11.0
111–115	39	0.39	39.0
116–120	7	0.07	7.0
121–125	1	0.01	1.0
126–130	1	0.01	1.0
131–135	3	0.03	3.0
136–140	1	0.01	1.0
Total	100	1.00	100.0

Table 2 Relative Frequency of Occupational Attainment

<i>Occupation</i>	<i>Frequency</i>	<i>Relative Frequency</i>	<i>Percentage</i>
Professional	33	0.33	33.0
Blue-collar	21	0.21	21.0
Craft	25	0.25	25.0
White-collar	12	0.12	12.0
Menial	9	0.09	9.0
Total	100	1.00	100.0

complicated data by obtaining the *proportion* of observations that fall in an interval (for QUANTITATIVE VARIABLES) or belong to a specific category (for CATEGORICAL variables).

The relative frequency for an interval (or a category) equals the number of observations in that interval (or category) divided by the total number of observations. A listing of intervals or categories and their relative frequencies is called a *relative frequency distribution*. Presenting the relative frequency distribution of a variable gives an overall picture of the distribution of data *if* the way in which data are grouped into different categories or intervals is appropriate. Table 1 is an example of a relative frequency distribution for a quantitative variable, intelligence quotient (IQ). Table 2 is an example of a relative frequency distribution for a QUALITATIVE VARIABLE, occupational attainment.

The relative frequency for the first interval in Table 1 is $1/100 = .01$. That is, one person out of 100 people,

for a relative frequency of .01, has an IQ score between 81 and 85. Note that this hypothetical distribution arranges data into mutually exclusive intervals of five units. Likewise, the relative frequency for the first category in Table 2 is $33/100 = 0.33$. It might be interpreted as “33 people out of 100 people, for a relative frequency of 0.33, work in professional positions.”

Data analysts often report the percentage for an interval or a category instead of a relative frequency. The percentage for an interval or category is simply a relative frequency multiplied by 100. The far right column of both Table 1 and Table 2 reports the percentage for each classification or grouped interval. The total sum of relative frequencies should equal 1 because the frequencies are proportions. It is also clear that the total sum of percentages should equal 100. However, at times, the rounding process may lead to slightly different total sums, such as 0.998 (for relative frequency) or 100.01 (for percentage).

Note that the use of relative frequency or percentage implies considerable stability in data. Thus, if the total number of cases is fairly small, say, 15, then one should be cautious about using the relative frequency distribution. Hubert M. Blalock (1979) suggests two rules of thumb: (a) Always report the number of observations (i.e., frequency) along with relative frequencies and percentages, and (b) do not report a percentage or a relative frequency unless the total number of observations is in the neighborhood of 50 or more.

It is also noteworthy that the width of an interval for grouping quantitative data has a critical impact on how a relative frequency distribution will look. For example, regrouping data into intervals of 25 units for Table 1 may substantially change the shape of the relative frequency distribution and provides less information than the original Table 1.

However, there is no fixed answer for a question like, “Shall I use an interval of narrow width to get more detail, or shall I use a large interval size to get more condensation of the data?” It usually depends on the nature of the data and the purpose of the research. A convention is that 10 to 20 intervals give a good balance between condensation and detail. More information on reporting relative frequency may be available in Hays (1994) and Johnson and Bhattacharyya (2001).

Finally, data analysts can further use a HISTOGRAM to present graphically any sort of frequency distribution, regardless of the scale of measurement of the data. A histogram and its variants, such as BAR GRAPHS and PIE CHARTS, are useful ways to picture frequency

distributions. A good discussion of general principles for creating a histogram may be found in Cleveland (1993).

—Cheng-Lung Wang

REFERENCES

- Blalock, H. M. (1979). *Social statistics* (rev. 2nd ed.). New York: McGraw-Hill.
- Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart.
- Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt.
- Johnson, R. A., & Bhattacharyya, G. K. (2001). *Statistics: Principles and methods* (4th ed.). New York: Wiley.

RELATIVE FREQUENCY DISTRIBUTION

A RELATIVE FREQUENCY distribution is one of several types of graphs used to describe a relatively large set of data. Examples are given in Figures 1 and 2. The first is for a quantitative variable, height in inches of males in the age range 20–25. The second is for course grades represented by the letters A–F, as commonly used in the U.S. system. These graphs are referred to as “relative” frequency distributions because the y-axis is a proportion scale, not a frequency scale. If 20 of 200 students are in the grade class “A,” then the relative frequency of this class is .10.

To create these graphs, one begins by grouping the full set of data into intervals or classes. The intervals or classes are plotted on the *x*-axis of a graph, using the midpoint of intervals or the label for classes. Bars (or columns) extend vertically above the *x*-axis to the point on the *y*-axis that represents the relative frequency of the interval or class.

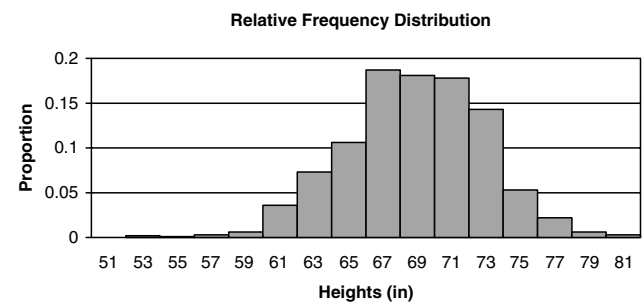


Figure 1 A Distribution of Male Heights Displayed According to the Relative Frequency for Each 2-Inch Interval Between 52 and 82 Inches

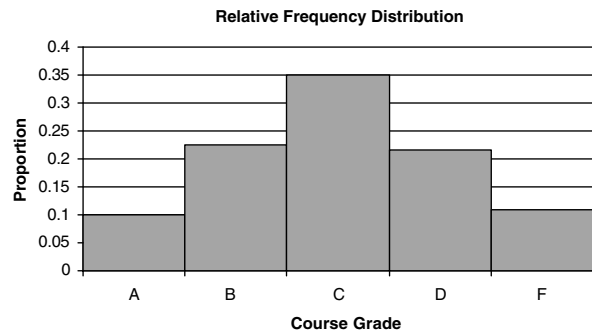


Figure 2 The Distribution of Course Grades Displayed According to Their Relative Frequencies

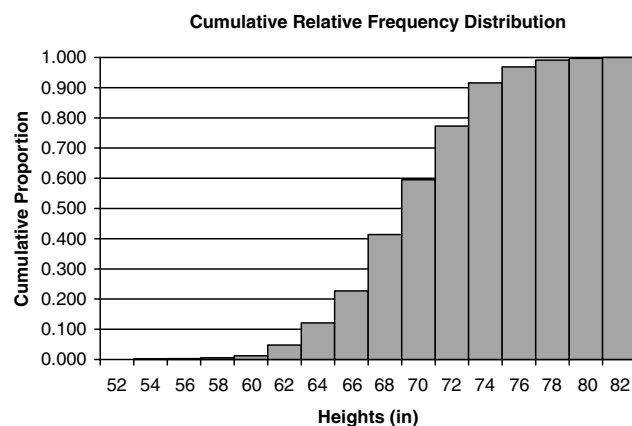


Figure 3 A Cumulative Distribution of the Height Data in Figure 1; the y-Axis Values $\times 100$ Are Percentiles

A variation on the relative frequency distribution is to graph cumulative proportions. Using this method, proportions multiplied by 100 become percentiles. Figure 3 is a cumulative display of the height data in the relative frequency distribution for heights. Note that the x -axis values are the upper limits of intervals, not their midpoints as before.

—Kenneth O. McGraw

RELATIVE VARIATION (MEASURE OF)

The coefficient of relative variation (CRV) is the STANDARD DEVIATION of a variable normed by dividing by its MEAN. This measure is used to give a sense of how large the standard deviation is. Yule and Kendall (1968, pp. 143–144) consider it to be an absolute measure of

dispersion because the division by the mean controls for the variable's unit of measurement.

The formula for the coefficient of relative variation is $CRV = s/\bar{X}$, where s is the standard deviation of the variable and $\bar{X}$ is its mean (see also VARIATION, COEFFICIENT OF). The statistic is appropriate only for a ratio VARIABLE, because shifting the scale of an interval variable would change its mean and therefore the CRV. Division by values near zero produce very large values, so the coefficient of relative variation is considered unreliable when the mean is near zero. A variant of the formula is using the absolute value of the mean in the denominator, so that CRV is always positive. The relative variation of observation i can be defined as the DEVIATION of the i th value x_i on variable X from its mean, divided by the mean: $RV_i = (X_i - \bar{X})/\bar{X}$. The coefficient of relative variation is then equivalent to the square root of the average of the squared relative variations:

$$CRV = \sqrt{\frac{\sum (RV_i)^2}{N}}.$$

The coefficient of relative variation is useful in comparing the variation on variables with different mean values. For example, consider comparing the variances of crime rates across the months of a year in cities with very different average amounts of crime. There might be much greater variance in crime between months in New York City than in Peoria, simply because the crime rate is so much larger in New York City. Dividing the standard deviations by their respective means gives a normed result that better reflects whether crime rates really are more variable across the months of a year in New York City than in Peoria.

This coefficient is also useful for comparing the variation of variables measured in different units. For example, is there more variability to people's age or their number of years of education? There is generally greater variance on age than on education for adults, but that might not be the case when the standard deviations are normed by dividing through by their means. For example, if the mean age of a set of adults is 50 with a standard deviation of 20, the relative variation on age is 0.40. If their mean number of years of education is 12 years with a standard deviation of 6, the relative variation on education is 0.50 (see Table 1). There is greater variance on age than education in this example, but education actually has greater relative variation once the means are taken into account.

—Herbert F. Weisberg

Table 1 Hypothetical Age and Education Data

<i>Person</i>	<i>Age</i>	<i>Years of Education</i>
1	20	15
2	30	21
3	40	18
4	50	12
5	60	9
6	70	6
7	80	3
Mean	50	12
Standard Deviation	20	6
Coeff. Relative Variation	0.40	0.50

REFERENCES

- Blalock, H. M. (1979). *Social statistics* (rev. 2nd ed.). New York: McGraw-Hill.
- Snedecor, G. W. (1956). *Statistical methods*. Ames: Iowa State College Press.
- Weisberg, H. F. (1992). *Central tendency and variability*. Newbury Park, CA: Sage.
- Yule, G. U., & Kendall, M. G. (1968). *An introduction to the theory of statistics*. New York: Hafner.

RELATIVISM

Relativism is a stance of systematic and thoroughgoing doubt about OBJECTIVIST, ESSENTIALIST, and foundational positions. It is not so much a position as an antiposition or a metaposition. Relativist arguments emphasize the inescapable role of RHETORIC in constructing claims as objective and foundational, as well as the contingency of tests, criteria, rules, experimentation, and other procedures that are claimed to guarantee objective foundations. Relativists do not deny the possibility of judgment (as is often claimed), but stress that judgment is bound up with broader human values, cultures, ways of life, and political visions in a way that cannot reduce to objective standards. Relativism is related to broader theoretical developments in social science; it can be seen as a philosophical correlate of POSTMODERNISM and constructionism, and the turn to discourse.

Relativism has been one of the most controversial, and often most despised, positions in social science. To say an approach or form of analysis has "slipped into relativism" was often considered sufficient to dismiss it. To be a relativist has been considered tantamount to endorsing (or at least not criticizing) the Holocaust or

spousal abuse, to be taking a position from which no judgment is possible and "anything goes."

Recently, debates about relativism often focus on the relation between REALISM and political or ethical critique. Some realists have argued that relativism fails to have sufficient purchase to develop political claims. Others have mobilized "death and furniture" arguments that highlight familiar and shocking atrocities (the massacre of the fleeing Iraqi army on the Basra road in 1991, female circumcision) or involve a sharp tap on an available table (Edwards, Ashmore, & Potter, 1995). Similar debates have emerged in feminism, where different ideas about gender and the nature of truth and politics have become contentious (Hepburn, 2000).

Richard Rorty (1991) has rethought issues in epistemology, ethics, and political philosophy along more relativist lines. He drew on Jacques Derrida's deconstructive philosophy to highlight the role of writing in constructing versions, making judgments, and constructing positions as apparently solid and timeless. His antifoundationalist arguments have considered the way judgments are grounded in claims that are themselves socially contingent. Barbara Herrnstein Smith (1997) argued that politics, morality, and cultural appraisal would be enriched rather than damaged by the relativist recognition of the failures of truth, rationality, and other notions central to objectivism.

Harry Collins (1985) has argued that methodological relativism is essential for a coherent study of scientific knowledge. He was a key figure in a sweep of new sociological work on science. This was distinct in not taking current scientific orthodoxy in any field as its start point, however plausible or established it might appear, because to do so would end up generating asymmetrical analyses (treating the orthodoxy as warranted by scientific observation, whereas the alternatives require social explanation). For example, in his research on gravitational radiation, Collins studied the rhetorical and persuasive work through which certain experiments were constructed as competent replications and certain claims were socially produced as incredible. Instead of starting with the then-conventional view that gravity waves had been found not to exist, he attempted to treat pro- and anti-gravity wave researchers in the same way. For this kind of analysis, instead of viewing scientific controversies as determined by a set of socially neutral rules and criteria that stand outside the controversy, criteria such as replication and testability become features of the controversy, with different scientists making different

judgments about what replication to trust and what test is definitive.

The centrality of relativism in the sociology of scientific knowledge is one of the reasons for its wider importance in contemporary social science. Building on this work, Malcolm Ashmore (1989) has developed a reflexive position that presses for an even more thoroughgoing relativism than Collins's. The need for methodological relativism in the sociology of scientific knowledge is now widely accepted. Jonathan Potter (1996) has argued that similar considerations are required when analyzing "factual accounts" in realms outside of science; otherwise, the same issues of taking sides and reinforcing the epistemic status quo will return.

Heated debate continues about moral and political consequences of relativist thinking.

—Jonathan Potter

REFERENCES

- Ashmore, M. (1989). *The reflexive thesis: Wrighting sociology of scientific knowledge*. Chicago: University of Chicago Press.
- Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice*. London: Sage.
- Edwards, D., Ashmore, M., & Potter, J. (1995). Death and furniture: The rhetoric, politics, and theology of bottom line arguments against relativism. *History of the Human Sciences*, 8, 25–49.
- Hepburn, A. (2000). On the alleged incompatibility between feminism and relativism. *Feminism and Psychology*, 10, 91–106.
- Potter, J. (1996). *Representing reality: Discourse, rhetoric and social construction*. London: Sage.
- Rorty, R. (1991). *Objectivity, relativism, and truth: Philosophical papers volume 1*. Cambridge, UK: Cambridge University Press.
- Smith, B. H. (1997). *Belief and resistance: Dynamics of contemporary intellectual controversy*. Cambridge, MA: Harvard University Press.

RELIABILITY

Theories of reliability have been primarily developed in the context of psychology-related disciplines since Charles Spearman laid the foundation in 1904. Nevertheless, the emphasis on reliability in these areas does not imply that this issue is of little relevance to other disciplines. Reliability informs researchers

about the relative amount of random inconsistency or unsystematic fluctuation of individual responses on a measure. The reliability of a measure is a necessary piece of evidence to support the CONSTRUCT VALIDITY of the measure, although a reliable measure is not necessarily a valid measure. Hence, the reliability of any measure sets an upper limit on the construct validity.

An individual's response on a measure tends to unsystematically vary across circumstances. These unsystematic variations, referred to as RANDOM ERRORS or random errors of MEASUREMENT, create inconsistent responses by a respondent. Random errors of measurement result from a variety of factors that are unpredictable and do not possess systematic patterns. These factors include test-takers' psychological states and physiological conditions, characteristics of the test environments, different occasions of test administration, characteristics of the test administrators, or different samples of test items in which some are ambiguous or poorly worded. As a result, one or more of the above factors contributes to the unpredictable fluctuation of responses by a person should the individual be measured several times.

TYPES OF MEASUREMENT ERRORS

It is safe to say that most, if not all, measures are imperfect. According to a recent development in measurement theory, the scores for any measured variable likely consist of three distinguishable components: the construct of interest, constructs of disinterest, and random errors of measurement (Judd & McClelland, 1998). Of these, the latter two components are measurement errors, which are not desirable but are almost impossible to eliminate in practice. The part of the score that represents constructs of disinterest is considered SYSTEMATIC ERROR. This type of error produces orderly but irrelevant variations in respondents' scores. In contrast, a measure's random errors of measurement do not impose any regular pattern on respondents' scores. It should be noted that errors of measurement in a score are not necessarily a completely random event for an individual (e.g., Student Smith's test anxiety); however, across a very large number of respondents, the causes of errors of measurement likely vary randomly and act as RANDOM VARIABLES. Hence, errors of measurement are assumed to be random. It is the random errors of measurement that theories of reliability attempt to understand, quantify, and control.

EFFECTS OF RANDOM ERRORS ON A MEASURE

Inconsistent responses on measures have important implications for scientific inquiries as well as practical applications. Ghiselli, Campbell, and Zedeck (1981) pointed out that inconsistent responses toward a measure have little value if researchers plan to compare two or more individuals on the same measure, place a person in a group based on his or her ability, predict one's behavior, assess the effects of an intervention, or even develop theories. Suppose a university admissions officer wants to select one of two applicants based on an entrance examination. If applicants' responses on this test tend to fluctuate 20 points from one occasion to another, an actual 10-point difference between them may be attributed to errors of measurement. The conclusion that one applicant has a stronger potential to succeed in college than another could be misleading. Similarly, a decrease of 10 points on an anger measure after a person receives anger management treatment may result from unreliable responses on the measure. The amount of errors of measurement resulting from one or more factors can be gauged by a variety of reliability coefficients, which will be discussed in a later section.

CLASSICAL TRUE SCORE THEORY

There have been two major theories of reliability developed since 1904: classical true score theory, based on the model of parallel tests, and GENERALIZABILITY THEORY, based on the model of domain sampling. Historically, Spearman (1904, 1910) started to develop classical TRUE SCORE theory, and it was further elaborated on by Thurstone (1931), Gulliksen (1950), and Guilford (1954). Generalizability theory is a newer theory, proposed by Jackson (1939) and Tryon (1957), and later advanced by Lord and Novick (1969) as well as Cronbach and his colleagues (1972). Conventional reliability coefficients such as parallel-forms and alternate-forms reliability, TEST-RETEST RELIABILITY, INTERNAL consistency RELIABILITY, SPLIT-HALF RELIABILITY, INTERRATER RELIABILITY, and INTRA-CODER RELIABILITY are estimated based on classical true score theory, although these coefficients can also be estimated by generalizability theory (see a demonstration in Judd & McClelland, 1998). Because generalizability theory and its applications are described elsewhere in the *Encyclopedia*, we will focus only on classical true score theory in the remaining section.

Classical true score theory is a simple, useful, and widely used theory to describe and quantify errors of measurement. In general, classical theory operates under the following assumptions:

1. $X_i = T + E_i$.
2. $E(X_i) = T$.
3. $F(X_1) = F(X_2) = F(X_3) = \dots = F(X_k)$.
4. $\rho_{X_1 X_2} = \rho_{X_1 X_3} = \rho_{X_1 X_k} = \dots = \rho_{X_{(k-1)} X_k}$.
5. $\rho_{X_1 Z} = \rho_{X_2 Z} = \rho_{X_3 Z} = \dots = \rho_{X_k Z}$.
6. $\rho_{E_1 E_2} = \rho_{E_1 E_3} = \rho_{E_1 E_k} = \dots = \rho_{E_{(k-1)} E_k} = 0$.
7. $\rho_{E_1 T} = \rho_{E_2 T} = \dots = \rho_{E_k T} = 0$.

Assumption 1, $X_i = T + E_i$, states that an observed score (or fallible score) of a test i (X_i), consists of two components: true score (T) and random errors of measurement, E_i . Errors of measurement, caused by a combination of factors described earlier, can be either positive or negative across different parallel tests (defined later), and therefore make the observed scores either higher or lower than the true score.

A true score on a test does not represent the true level of the characteristic that a person possesses. Instead, it represents the *consistent* response toward the test, resulting from a combination of factors. Note that different combinations of factors can lead to a variety of consistent responses. For instance, consistent responses on a measure of intelligence in an environment with humidity, noise, and frequent interruptions are different from those on the same measure under a cool, quiet, and controlled environment. The former consistent response represents the true score for intelligence under the *hardship* environment, whereas the latter response is the true score for intelligence under the *pleasant* condition.

Assumption 2, $E(X_i) = T$, states that the true score is the expected value (mean) of an individual's observed scores, should the person be tested over an infinite number of parallel tests. For instance, assuming an individual receives a large number of k parallel tests, the observed score of each test for this person can be expressed as

$$X_1 = T + E_1,$$

$$X_2 = T + E_2,$$

$$\vdots$$

$$X_k = T + E_k.$$

Because errors of measurement are either positive or negative, the sum of these errors in the long run is

zero, $E(E_i) = 0$. As a result, the expected value of observed scores equals the true score.

Assumptions 3 through 5 are about the characteristics of parallel tests, which have theoretical as well as practical implications. Assumption 3, $F(X_1) = F(X_2) = F(X_3) = \dots = F(X_k)$, states that distributions of observed scores across k parallel tests are identical for a large population of examinees. According to the assumption of identical distributions, we expect $T_1 = T_2 = \dots = T_k$ and $\sigma_{X_1}^2 = \sigma_{X_2}^2 = \dots = \sigma_{X_k}^2$. In addition, CORRELATIONS among observed scores of the k parallel tests are the same, which is expressed as $\rho_{X_1X_2} = \rho_{X_1X_3} = \rho_{X_1X_k} = \dots = \rho_{X_{(k-1)}X_k}$ (Assumption 4). Finally, Assumption 5, $\rho_{X_1Z} = \rho_{X_2Z} = \rho_{X_3Z} = \dots = \rho_{X_kZ}$, states that there are equal correlations between observed scores and any other measure (Z), which is different from k parallel tests.

Assumptions 6 and 7 essentially state that errors of measurement are completely random. Consequently, correlations among errors of measurement across k parallel tests in a large POPULATION of examinees equal zero, as do the correlations between errors of measurement and the true score. Based on Assumption 7, a VARIANCE relationship among σ_X^2 , σ_T^2 , and σ_E^2 can further be expressed by $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$. Specifically, the observed score variance in a population of examinees equals the sum of the true score variance and the errors of measurement variance. Deduced from the above relationship in conjunction with Assumption 3, errors of measurement variances across k parallel tests are equal to one another, $\sigma_{E_1}^2 = \sigma_{E_2}^2 = \dots = \sigma_{E_k}^2$.

THEORETICAL DEFINITION OF RELIABILITY

According to the variance relationship $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$, a perfectly reliable measure occurs only when random errors do not exist (i.e., $\sigma_E^2 = 0$ or $\sigma_X^2 = \sigma_T^2$). When σ_E^2 increases, observed differences on a measure reflect both true score differences among the respondents and random errors. When $\sigma_X^2 = \sigma_E^2$, all of the observed differences reflect only random errors of measurement, which suggests that responses toward the measure are completely random and unreliable.

If we divide σ_X^2 on both sides of the above equation, the variance relationship can be modified as

$$1 = \frac{\sigma_T^2}{\sigma_X^2} + \frac{\sigma_E^2}{\sigma_X^2}.$$

Reliability of a measure is defined as

$$\frac{\sigma_T^2}{\sigma_X^2},$$

which can be interpreted as the squared correlation between observed and true scores (ρ_{XT}^2). Conceptually, reliability is interpreted as the proportion of the observed score variance that is explained or accounted for by the true score variance. Following the above assumptions, it further proves that a correlation between two parallel tests equals

$$\frac{\sigma_T^2}{\sigma_X^2}.$$

That is,

$$\rho_{X_1X_2} = \frac{\sigma_{T_1}^2}{\sigma_{X_1}^2} = \frac{\sigma_{T_2}^2}{\sigma_{X_2}^2}.$$

If a correlation between two parallel tests is 0.80, it suggests that 80% (not 64%) of the observed score variance is accounted for by the true score variance. We will hereafter use the general notation, $\rho_{XX'}$ and $r_{XX'}$, to represent reliability for a population and a SAMPLE, respectively.

PROCEDURES TO ESTIMATE RELIABILITY

A variety of reliability coefficients are based on classical true score theory, such as test-retest reliability, internal consistency reliability, interrater reliability, or intracoder reliability, which are reviewed elsewhere in the *Encyclopedia*. We will focus on three additional reliability coefficients: alternate-forms reliability, reliability of difference scores, and reliability of composite scores.

An alternate-forms reliability coefficient, also referred to as coefficient of equivalence, is estimated by computing the correlation between the observed scores for two similar forms that have similar means, STANDARD DEVIATIONS, and correlations with other variables. Generally, a group of respondents receives two forms in a random order within a very short period of time, although the time period between administrations may be longer.

When conducting interventions or assessing psychological/biological development, researchers are often interested in any change in scores on the pre- and posttests. The difference between scores on the same or a comparable measure at two different times

is often referred to as a change score or difference score (denoted as D). The reliability of the difference scores in a sample can be estimated by

$$r_{DD'} = \frac{s_X^2 r_{XX'} + s_Y^2 r_{YY'} - 2s_X s_Y r_{XY}}{s_X^2 + s_Y^2 - 2s_X s_Y r_{XY}},$$

where s_X^2 and s_Y^2 are variances of the pre- and posttests, $r_{XX'}$ and $r_{YY'}$ are the reliabilities of the tests, and r_{XY} is the correlation between the tests. If the variances of both measures are the same, the above equation can be rewritten as

$$r_{DD'} = \frac{\frac{r_{XX'} + r_{YY'}}{2} - r_{XY}}{1 - r_{XY}}.$$

Given that $r_{XX'}$ and $r_{YY'}$ are constant, stronger r_{XY} results in weaker $r_{DD'}$. Furthermore, $r_{DD'}$ moves toward zero as r_{XY} approaches the average reliability of the measures. In reality, the size of r_{XY} is often strong because of two measures' similarity. As a result, difference scores tend to have low reliability, even when both measures are relatively reliable.

In contrast to difference scores, researchers sometimes need to utilize a composite score based on a linear combination of raw scores from k measures. For example, a personnel psychologist may select job applicants based on a composite score of an intelligence test, a psychomotor ability test, an interest inventory, and a personality measure. The reliability of this composite score (C) can be estimated by

$$r_{CC'} = 1 - \frac{\sum_{i=1}^k s_i^2 - \sum_{i=1}^k (r_{ii'} s_i^2)}{s_C^2},$$

where s_C^2 and s_i^2 are variances of the composite score, and $r_{ii'}$ is the reliability of measure i . In practice, researchers may create the composite score based on a linear combination of STANDARDIZED SCORES from k measures. The reliability can be estimated by the simplified formula,

$$r_{CC'} = 1 - \frac{k - k\bar{r}_{XX'}}{k + (k^2 - k)\bar{r}_{XY}},$$

where $\bar{r}_{XX'}$ is the average reliability across k measures, and $\bar{r}_{XY}$ is the average intercorrelation among the tests. Assuming the average intercorrelation is constant, the reliability of the composite score is greater than the average reliability of the measures. Hence, the reliability of a composite score tends to be relatively more reliable.

STANDARD ERRORS OF MEASUREMENT

As noted earlier, reliability coefficients only inform about the *relative* amount of random inconsistency of individuals' responses on a measure. A test with a Cronbach's alpha of .90 suggests that the test is more internally consistent than another test with a Cronbach's alpha of .70. However, both reliability estimates do not provide *absolute* indications regarding the precision of the test scores. For instance, the Stanford-Binet Intelligence Scale is a very reliable measure with an internal consistency reliability of .95 to .99. Although the test is highly reliable, we are not confident in saying that a score of 60 is truly higher than the average of 50 by 10 points. More realistically, the 10 points reflect the true difference in intelligence *and* random errors of measurement.

In order to interpret test scores accurately, researchers need to take unreliability into consideration, which is gauged by the square root of the errors of measurement variance, or standard error of measurement (σ_E). Standard error of measurement, derived from

$$1 = \frac{\sigma_T^2}{\sigma_X^2} + \frac{\sigma_E^2}{\sigma_X^2},$$

is expressed as $\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}$ in a population, which can be estimated by $\hat{\sigma}_E = s_E = s_X \sqrt{1 - r_{XX'}}$ from a sample. Conceptually, the standard error of measurement reveals how much a test score is expected to vary given the amount of error of measurement for the test. When σ_E is available, researchers can use it to form CONFIDENCE INTERVALS of the observed scores. Specifically, the observed score (X) in a sample has approximately 68%, 95%, and 99% confidence to fall in the range of $X \pm 1\hat{\sigma}_E$, $X \pm 2\hat{\sigma}_E$, and $X \pm 3\hat{\sigma}_E$, respectively.

FACTORS THAT AFFECT RELIABILITY ESTIMATES

While estimating the reliability of a measure, researchers cannot separate the respondents' characteristics from the measure's characteristics. For instance, test performance on an algebra test in part depends on whether the test items are difficult or easy. At the same time, a determination of the test items as easy or difficult relies on the ability of the test taker (e.g., college student vs. junior high school student; homogeneous group vs. heterogeneous group). When a group is so homogeneous that members

possess similar or the same algebraic ability, the variance of the observed scores could be restricted, which leads to a very low reliability even though the test could be perfectly reliable. The above examples clearly demonstrate that reliability estimates based on classical true score theory change contingent upon the group of test takers and/or the test items. Specifically, traditional procedures to estimate reliability are group dependent and test dependent, which suggests that the standard error of measurement (and reliability) of a measure is not the same for all levels of respondents. Empirical evidence has shown that the standard error of measurement tends to be smaller for the extreme scores than the moderate scores.

There are other characteristics of measures that affect reliability. First, according to the Spearman-Brown Prophecy formula, the reliability of a measure tends to increase when the length of the measure increases, given that appropriate items are included. The Spearman-Brown Prophecy formula is defined as

$$\hat{\rho}_{XX'} = \frac{n \times \rho_{XX'}}{1 + (n - 1)\rho_{XX'}},$$

where n is the factor by which an original test is lengthened, and $\hat{\rho}_{XX'}$ and $\rho_{XX'}$ are reliability estimates for the lengthened test and the original test, respectively. Second, empirical findings have shown that reliability coefficients tend to increase when response categories of a measure increase. A MONTE CARLO study has revealed that overall reliability estimates of a 10-item measure, such as test-retest reliability or CRONBACH'S ALPHA, increase when response categories increase from a 2-point scale to a 5-point scale. After that, the reliability estimates become stable. Test content can also affect reliability. For instance, the reliability of power tests can be artificially inflated if time limits are not long enough. Power tests refer to tests that examine respondents' knowledge of subject matter. When time limits are short, most respondents cannot answer all the items. As a result, their responses across items become artificially consistent.

Finally, reliability can be affected by the procedures researchers opt to use. As described earlier, reliability coefficients based on classical true score theory estimate one or two types of errors of measurement. For instance, internal consistency estimates response variations attributed to item heterogeneity or inconsistency of test content, whereas test-retest reliability estimates variations attributed to changes over time, including carry-over effects. In contrast, alternate-forms

reliability with a long lapse of time (e.g., 4 months) estimates two types of random errors, namely, those attributed to changes over time and inconsistency of test content. Although all reliability coefficients based on classical true score theory estimate the relative amount of errors of measurement, the errors estimated by different reliability coefficients are *qualitatively* different from each other. Hence, it is not logical to compare two different types of reliability estimates.

The fact that different random errors are being estimated by different reliability coefficients has an important implication when researchers *correct* for attenuation due to unreliability. Derived from the assumptions of classical true score theory, a population correlation between two true scores ($\rho_{T_X T_Y}$) can be adversely affected by unreliability of the measures (X and Y). The adverse effect can be deduced from the relationship between a population correlation and population reliabilities of measures X and Y ($\rho_{XX'}$ and $\rho_{YY'}$), which is defined as

$$\rho_{T_X T_Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'} \rho_{YY'}}}.$$

Often, researchers want to show what the sample correlation ($\hat{r}_{XY}$) would be if both measures were perfectly reliable. Following the modified formula from above, the disattenuated correlation can be obtained by

$$\hat{r}_{XY} = \frac{r_{XY}}{\sqrt{r_{XX'} r_{YY'}}}.$$

The difference between $\hat{r}_{XY}$ and r_{XY} increases when one or both measures become less reliable. For instance, $\hat{r}_{XY}$ equals r_{XY} when $r_{XX'} = r_{YY'} = 1$. However, the size of $\hat{r}_{XY}$ becomes twice as much as r_{XY} when $r_{XX'} = r_{YY'} = 0.50$. In practice, the correction for attenuation tends to overestimate the actual correlation. Foremost, it does not seem tenable to assume that both $r_{XX'}$ and $r_{YY'}$ deal with the *same* type(s) of errors of measurement. Furthermore, there are many different types of errors of measurement that are not adequately estimated by classical true score theory. For instance, researchers may be interested in finding out how reliable an achievement test is when it is administered by both male and female administrators at two different times and at three different locations.

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

- Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. New York: Harcourt Brace Jovanovich.

- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.
- Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: W. H. Freeman.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1). Boston: McGraw-Hill.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

RELIABILITY AND VALIDITY IN QUALITATIVE RESEARCH

Among QUALITATIVE RESEARCHERS, the concepts of RELIABILITY and VALIDITY have eluded singular or fixed definitions. As the understanding of qualitative inquiry has evolved in varied ways, especially over the past 20 or so years, so have the understandings of both concepts.

RELIABILITY

In its original formulation among quantitative researchers, a study was considered reliable if it could be replicated by other researchers. Among qualitative researchers, there is a considerable division of thinking as to whether replication or the possibility of reproducing results has any import whatsoever in terms of judging the value of qualitative studies. Some people, such as Jerome Kirk and Marc Millar (1986), argue that reliability in the sense of repeatability of observations still has an important epistemic role to play in qualitative inquiry. They contend that for a study to be judged good or valid, the observations made in that study must be stable over time, and that different methods, such as interviews and observations, should yield similar results.

Others, most particularly Egon Guba and Yvonna Lincoln (1994), have noted that because of the philosophic assumptions underlying qualitative inquiry (e.g., reality is constructed), the concept of reliability should give way to the analog idea of dependability. They contend that replication is not possible for qualitative inquiry; all that can or should be expected from a researcher is a careful account of how she or he obtained and analyzed the data.

Finally, some now hold that there is little point in continuing to pose reliability as a criterion for judging the quality of studies. Although it is possible that one researcher in a particular setting may well repeat the observations and findings of another researcher in that setting, one should not be concerned if this does not happen. The basis for this claim derives from the realization that theory-free observation is impossible, or that researchers can observe the world only from a particular place in that world (Smith & Deemer, 2000). If this is so, then reliability can be given no special epistemic status and need not be an issue of consequence for qualitative researchers.

VALIDITY

For QUANTITATIVE RESEARCHERS, validity is a concept of major epistemic import. To say that a study is valid is to say that it is a good study in that the researcher has accurately represented the phenomena or reality under consideration. For various related reasons, most particularly the rejection of naive or direct REALISM and the rejection of an epistemological foundationism, qualitative researchers do not find this definition of validity acceptable. Qualitative researchers have replaced realism and foundationalism with a variety of positions ranging from indirect or subtle realism and epistemological quasifoundationalism (fallibilism) to nonrealism and nonfoundationalism with numerous variations on theme in between (Smith, 1993). Because of this situation, the efforts by qualitative researchers to replace the traditional understanding of validity has led to an almost bewildering array of definitions and variations on definitions for this concept.

Martyn Hammersley (1990) is representative of those who have advanced an understanding of validity based on a subtle or indirect realism, combined with a quasifoundationalism or fallibilist epistemology. For him, a judgment of validity is not a matter of the extent to which an account accurately reproduces reality, but rather a judgment about the extent to which an account is faithful to the particular situation under consideration. This means that even though we have no direct access to reality and our judgments are always fallible, there still can and must be good reasons offered for claiming an account as true. These considerations lead to a definition of a valid study as one whose results have met the tests of plausibility and credibility. The former is a matter of whether or not an account of a situation is likely true given the existing state of knowledge of that situation. The latter directs attention to whether or not

a researcher's judgment is accurate given the nature of the phenomena, the circumstances of the research, the characteristics of the researcher, and so on.

A second broad approach, which also is, to one degree or another, based on an indirect realism and a quasifoundationalism, involves a wide variety of what can be called "hyphenated" definitions of validity. Numerous terms, such as *catalytic*, *situated*, *voluptuous*, *transgressive*, *ironic*, and *interrogated*, have been placed as modifiers to the term *validity*. Both David Altheide and John Johnson (1994) and Patti Lather (1993) have discussed these and various other possibilities.

Finally, some qualitative researchers see little reason to bother with the concept of validity. Because these people have accepted nonrealism and nonfoundationalism, they see little point in continuing to talk about true accounts of the world or of accurate depictions of an independently existing reality. They claim that all that is available to us are different linguistically mediated social constructions of reality. As such, the most that can be said for validity is that it is a matter of social agreement, and such agreements are not epistemological, but rather political and moral ones (Schwandt, 1996; Smith & Deemer, 2000).

—John K. Smith

REFERENCES

- Altheide, D., & Johnson, J. (1994). Criteria for assessing interpretive validity in qualitative research. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (1st ed., pp. 485–499). Thousand Oaks, CA: Sage.
- Guba, E., & Lincoln, Y. (1994). Competing paradigms in qualitative research. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (1st ed., pp. 105–117). Thousand Oaks, CA: Sage.
- Hammersley, M. (1990). *Reading ethnographic research*. London: Longman.
- Kirk, J., & Millar, M. (1986). *Reliability and validity in qualitative research*. Beverly Hills, CA: Sage.
- Lather, P. (1993). Fertile obsession: Validity after post-structuralism. *Sociological Quarterly*, 35, 673–694.
- Schwandt, T. (1996). Farewell to criteriology. *Qualitative Inquiry*, 2, 58–72.
- Smith, J. (1993). *After the demise of empiricism: The problem of judging social and educational inquiry*. Norwood, NJ: Ablex.
- Smith, J., & Deemer, D. (2000). The problem of criteria in the age of relativism. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 877–922). Thousand Oaks, CA: Sage.

RELIABILITY COEFFICIENT

A reliability coefficient is a measure of the proportion of true variance relative to the total observed score variance resulting from the application of a measurement protocol (such as a test, a scale, an expert rating, etc.) to a population of individuals.

Any measurement operation of a given attribute provides an observed score for that attribute, which, in principle, may be broken down into a TRUE SCORE component plus a measurement error component. The true score is conceived as an unknown constant within each individual, whereas the error component is modeled as a random variable uncorrelated with both the true score and other error components. Based on these properties, it has been algebraically demonstrated that the squared correlation of *observed* and *true* scores for a given population equals the proportion of the *true score* variance relative to the *observed score* variance, which is a reliability coefficient (see RELIABILITY). Thus, for a perfectly reliable assessment, the observed score variance is entirely accounted for by the true variance, and the resulting coefficient is 1.00. Conversely, for a totally unreliable assessment, the observed score variance has nothing to do with actual differences in the attribute to be measured, and the resulting reliability coefficient is .00. Although no absolute standard for a reliability coefficient exists, .70 and .90 are considered as the minimum acceptable values of reliability for group comparison and for making individual decisions, respectively.

So far, we defined a reliability coefficient in terms of observed score and true score variability. However, whereas the observed score is available after any measurement operation, the true score is, by definition, impossible to get to directly. In spite of this, the reliability coefficient can be obtained indirectly by looking at the consistency of the observed scores across repeated measures. The most popular methods for examining score consistency are (a) TEST-RETEST RELIABILITY, (b) alternate-form (or equivalent-form) reliability, (c) SPLIT-HALF RELIABILITY, and (d) CRONBACH'S ALPHA coefficient.

Methods (a) and (b) address the question of whether a test, or statistically equivalent parallel forms of a test, given twice to the same individuals, would provide the same set of observed scores on each occasion. Methods (c) and (d) address the question of whether statistically equivalent parallel empirical indicators used to

measure a single theoretical concept would correlate with each other on a single occasion.

Some research designs involve behavior rating by experts according to a predefined scale or coding procedure. INTERRATER RELIABILITY and intrarater reliability coefficients represent additional types of reliability coefficients that apply to these measurement protocols.

Once the reliability coefficient has been evaluated by any of the above methods, one can use it to compute the standard error of measurement, which can be used to estimate—punctually or from a CONFIDENCE INTERVAL—one's true score by one's observed score.

Factors influencing the reliability coefficient are test length (i.e., tests comprised of more items provide higher reliability coefficients), examinee heterogeneity (i.e., reliability increases as individual differences in the attribute to be measured increase), and test difficulty (i.e., a moderately difficult test increases the likelihood of detecting individual differences).

—Marco Lauriola

REFERENCES

- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

REPEATED MEASURES

The term *repeated measures* refers to how the design for the data collection is such that data are collected more than once on each unit. The unit may be a child, with the pediatrician measuring the weight of the child at yearly intervals; a country, with measures of inflation over several years; or a couple, with incomes for husband and wife. Repeated measures often involve data collected over time. The problem now is that there is no independence between the observations for a unit the way there is when there is only one observation on each of many units.

Repeated measures can involve data collected only twice. This is the paired data design, and it leads to the *T-TEST* for paired data. Husband-wife income data are of this kind, as are experimental data with a before-after design. It is tempting to simply do a

t-test for the difference between two gender means, but that does not take into account the paired nature of the data. The numerators for the *t*-test for the difference between two means and the paired *t*-test are identical, and equal to the difference between the means for the two groups. The difference is in the denominators. Suppose one couple has high incomes and another couple has low incomes. This variation is included in the STANDARD DEVIATION and therefore increases the denominator in the test for two means, making for a smaller value of *t* and therefore a larger *P* VALUE. In the paired test, what matters are the *differences* between husbands' and wives' incomes, and it does not take into account whether the original values are high or low. By eliminating this source of variation, the denominator becomes smaller, the *t* value becomes larger, and the corresponding *p* value becomes smaller and more significant. The only adjustment is in the DEGREES OF FREEDOM. The test for the difference of two means has $n_1 + n_2 - 2$ degrees of freedom, whereas the paired test has only $n - 1$ degrees of freedom.

The paired *t*-test is equivalent to a two-way ANALYSIS OF VARIANCE, with one observation in each cell and the interaction mean square used in the denominator for the *F*s instead of the residual mean square. One factor is gender, with 2 values, and the other factor is couple, with *n* values. Now we get the effect of gender after taking out the effect of couple. Numerically, *F* on 1 and $n - 1$ degrees of freedom for gender from the analysis of variance equals t^2 from the paired *t*-test.

If there is a factor with more than two categories, the paired *t*-test *generalizes* to a two-way analysis of variance with one observation in each cell, and the interaction mean square is used in the denominator for *F*. In this case, there is no corresponding *t*-test. Repeated measures analysis becomes TIME-SERIES analysis when the observations are made at known points in time on only one unit.

—Gudmund R. Iversen

REFERENCES

- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York, Springer-Verlag.
- Girden, E. R. (1992). *ANOVA: Repeated measures* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-084). Newbury Park, CA: Sage.
- Hays, W. L. (1988). *Statistics*. New York: Holt, Rinehart and Winston.

REPERTORY GRID TECHNIQUE

Derived from Kelly's personal construct psychology (see Fransella & Bannister, 1977), repertory grid technique is a method for representing and exploring a person's understanding of things. It is particularly useful for getting people to articulate their knowledge of topics to which they have previously given little thought, or find difficulty expressing. A simplified grid is shown in Table 1. It comprises four features:

- *A Focus.* In Table 1, the focus is the skills of team members in a manufacturing organization.
- *Elements.* These are the objects of thought. Elements are often people but can be places, ideas, or inanimate things.
- *Constructs.* These are the qualities the person uses to describe and differentiate between the elements. Constructs are bipolar and have positive and negative ends.
- *A Linking Mechanism.* This refers to how the elements are related to the constructs. It can be a categorical, binary, or continuous scale.

A common way of completing a grid is for the researcher to supply its focus. The participant is then asked to select examples of elements, for example, a skilled team member you know. Constructs are then

generated by "triading"—selecting three elements and stating how two are similar to, but different from, a third. The positive and negative sides of each construct are specified. Triading is repeated until no more constructs can be generated. The person then links the constructs to elements by scoring their relationship. Variations in grid completion include generating constructs by simply comparing and contrasting two elements (dyading), and supplying the elements and/or constructs. Some see supplying elements or constructs as counter to the aim of grids (i.e., to explore "personal" understanding).

Grids are analyzed by examining the meaning of constructs, as well as the relationships between constructs (e.g., do constructs cluster to form a superordinate factor?), between elements, and between constructs and elements (e.g., which constructs associate with which elements?). Qualitative and quantitative analytical procedures can be used separately or, preferably, in combination. The complexity of the grid means that quantitative techniques are particularly useful (see Leach, Freshwater, Aldridge, & Sunderland, 2001, for an overview). However, some argue that quantitative analysis detracts from or violates the interpretivist assumptions that underlie repertory grids, whereas others state that it can ignore the rich data generated from the interview itself. Qualitative analysis of construct meanings and the interview is therefore recommended.

Table 1 A Simplified Repertory Grid Showing a Manufacturing Manager's Understanding of His Team Members' Skills

Constructs (Positive Pole)	Elements						Constructs (Negative Pole)
	Skilled Operator	Skilled Operator	Averagely Skilled Operator	Averagely Skilled Operator	Unskilled Operator	Unskilled Operator	
	Leslie	Ann	Frank	Denise	Julie	David	
Time keeping	✓	✓	✓	×	✓	✓	Poor time keeping
Flexible	✓	✓	✓	✓	×	×	Unwilling to use skills
Quality awareness	×	×	✓	✓	×	×	Lacks awareness
Appropriate communication with manager	✓	✓	×	✓	×	×	Poor communication
Suggests ideas	✓	✓	✓	×	×	×	Doesn't want to improve performance
Competitive	✓	×	×	×	×	×	Uncompetitive
Solves problems on own	✓	✓	✓	✓	×	×	Always seeks help

SOURCE: Adapted from Easterby-Smith, Thorpe, and Holman (1996).

Finally, there are a number of limitations to the use of grids. It is a time-consuming process for all involved. The interview process can bring up personal and sensitive issues, and the interviewer needs to be skilled in dealing with these. Unless change is a specific focus, repertory grid interviews are not adept at elucidating change processes (i.e., how and why things change).

—David John Holman

REFERENCES

- Easterby-Smith, M., Thorpe, R., & Holman, D. (1996). Using repertory grids in management. *Journal of European Industrial Training*, 20, 1–30.
- Fransella, F., & Bannister, D. (1977). *A manual for repertory grid technique*. London: Academic Press.
- Leach, C., Freshwater, K., Aldridge, J., & Sunderland, J. (2001). Analysis of repertory grids in clinical practice. *British Journal of Clinical Psychology*, 40, 225–248.

REPLICATION

The purpose of a social science EXPERIMENT is to estimate (a) the effect of some treatment on certain output variables in members of a given POPULATION, (b) the difference(s) between two or more populations with respect to some variable(s), or (c) the relationship or CORRELATION between two or more variables within some specified population. All such experiments can lead to erroneous conclusions if the measures employed are invalid, if the samples are not representative of the specified populations, or if extraneous but unidentified factors influence the dependent variables. Therefore, before accepting the results of any such experiment as establishing an empirical fact, such experiments should be repeated or replicated.

The closest one might come to a *literal replication* might be for the original author to draw additional samples of the same size from the same populations, employing all the same techniques, measuring instruments, and methods of analysis as on the first occasion. The results of the original SIGNIFICANCE TESTING (e.g., “ $p \leq .01$ that Measures A and B are positively correlated in the population sampled”) predict the outcome of (only) a truly literal replication. That literal replications are seldom possible is no great loss inasmuch as they may succeed, not only if the original finding was

valid, but also if the finding was invalid but the author was consistent in making the same crucial mistakes.

Operational replication occurs when a new investigator repeats the first study, following as closely as possible the “experimental recipe” provided by the first author in the Methods section of his or her report. A successful operational replication provides some assurance that the first author’s findings were not the result of some unknown and uncontrolled variables.

A limitation of operational replication is that its success does not necessarily support the claim that, for example, “Constructs X and Y are positively correlated in Population Z,” but only that “Scores obtained using Measures A and B are positively correlated in samples drawn in a certain way from Population Z.” Those scores may not be good measures of the constructs, and/or the sampling methods may not produce samples representative of Population Z. Moreover, neither authors nor editors of the psychological literature have learned to limit the Methods sections of research reports to contain all of, and only, those procedural details that the first author considers to have been necessary and sufficient to yield the findings obtained.

The most risky but most valuable form of replication is *constructive replication*, in which the second author ignores everything in the first report except the specific empirical fact or facts claimed to have been demonstrated. The replicator chooses methods of sampling, measurement, and data analysis that he or she considers to be appropriate to the problem.

An experiment that claims to have confirmed a theoretical prediction can be usefully replicated in two ways: The validity of the prediction can be constructively replicated, or the incremental validity of the theory can be constructively replicated by experimentally testing new and independent predictions that follow from the theory. Thus, Lykken (1957, 1995) tested three predictions from the theory that the primary psychopath starts life at the low end of the BELL-SHAPED CURVE of genetic fearfulness. Several investigators constructively replicated the original findings, whereas Hare (1966), Patrick (1994), and others have experimentally confirmed new predictions from the same theory.

—David T. Lykken

REFERENCES

- Hare, R. D. (1966). Temporal gradient of fear arousal in psychopaths. *Journal of Abnormal and Social Psychology*, 70, 442–445.

- Lykken, D. T. (1957). A study of anxiety in the sociopathic personality. *Journal of Abnormal and Social Psychology*, 55, 6–10.
- Lykken, D. T. (1995). *The antisocial personalities*. Mahwah, NJ: Lawrence Erlbaum.
- Patrick, C. J. (1994). Emotion and psychopathy: Startling new insights. *Psychophysiology*, 31, 415–428.

REPLICATION/REPLICABILITY IN QUALITATIVE RESEARCH

This concept shares much with that of RELIABILITY in that both refer to the extent to which a research operation is consistently repeatable. If consistency between researchers is achieved, more faith is placed in the truth of the findings. REPLICATION refers to the extent to which a re-study repeats the findings of an initial study. Replicability is the extent to which such re-study is made feasible by the provision of sufficient information about research procedures in the first study.

These ideas depend on philosophical assumptions that are not shared by all QUALITATIVE RESEARCHERS. For example, researchers committed to an idealist vision of the way in which language constructs versions of reality often regard their own texts (research reports, etc.) as constructing their objects of study, rather than reporting truths in conventional realist fashion. If this relativist, social constructionist view is taken, it makes little sense to seek out replication, as each account is simply one more “partial truth” (Clifford & Marcus, 1986).

Where realist assumptions are made, replication in qualitative research is not without difficulties either. Erasure of the personal idiosyncracies of the researcher is fundamental to the natural scientific worldview and to some versions of quantitative social research, but this is well-nigh impossible in qualitative research. In fieldwork, a great deal of what emerges depends on the personal attributes, analytic insights, and social skills of the researcher, and these are unlikely to be shared fully by any second researcher. Much of the strength of qualitative research lies in in-depth examination of particular cases, seen in the richness of their (sometimes unique) contexts. Cases do not remain static over time and may not share very much with other cases that, at some other level, are judged to be the same. Observers looking at the “same” case at some later time, or attempting to study “similar” cases, may, in effect, be studying different things, and they

do so using different instruments (i.e., themselves). Attempts at replicating some of the more famous studies in qualitative social research (Mead and Freeman in Samoa; Lewis and Redfield in Mexico; Whyte and Boelen in Boston—summarized in Seale, 1999) have encountered these problems, and in retrospect, these attempts (and the embarrassment they provoked in their respective disciplines) appear naïve.

Replicability, however, is a more realistic goal, although perhaps it should be renamed “methodological accounting” in view of the practical difficulties reviewed. The documentation of research procedures, as well as reflection on the influence on a research project of prior assumptions and personal preferences, can help readers assess the adequacy with which researchers’ claims have been supported. Lincoln and Guba (1985) have outlined a scheme for AUDITING a research project, in which such documents and reflections are investigated by a team of research colleagues toward the end of a research project in order to assess methodological adequacy. Although such elaborate and expensive exercises are rarely carried out, in general, there has been a turn toward methodological accounting in qualitative research reports, and this is perhaps as far as is practical to go in the direction of the full experimental replications demanded by natural scientists as a matter of course.

—Clive Seale

REFERENCES

- Clifford, J., & Marcus, G. E. (Eds.). (1986). *Writing culture: The poetics and politics of ethnography*. Berkeley: University of California Press.
- Lincoln, Y. S., & Guba, E. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.
- Seale, C. F. (1999). *The quality of qualitative research*. London: Sage.

REPRESENTATION, CRISIS OF

The phrase *crisis of representation* refers to the claim that social researchers are unable to directly represent or depict the LIVED EXPERIENCES of the people they study. The implications of this claim are twofold. First, the phrase announces that the core ideas that had long dominated the understanding of researchers about the nature of social research are no longer viable. The classic norms of OBJECTIVITY, VALIDITY, truth as

correspondence, and so on are no longer applicable to our understanding of inquiry. Second, in response to the demise of these ideas, there is the further understanding that the lived experiences of people are not discovered by researchers to be depicted as they really are; rather, these experiences are constructed by researchers in their interpretations of the interpretations of others as presented in their written social texts. The implications of this shift from discovery or finding to constructing or making are still being worked out.

The phrase was, if not coined, certainly brought to the attention of many social researchers by George Marcus and Michael Fischer in 1986. However, prior to that, the idea that the direct or objective representation of the experiences of people by researchers was an impossibility had been intellectually fermenting over the course of the 20th century and even earlier. Tracings of this concern can be found, for example, in the late 19th- and early 20th-century antipositivist writings of various German scholars, in the realization at midcentury by many philosophers of science that there can be no possibility of theory-free observation, and in the acknowledgment in the 1970s and early 1980s that anthropological writings are interpretations of interpretations and that the boundaries of social science and the humanities had become blurred (Smith, 1989).

Today, the crisis of representation as a "crisis" has more historical than contemporary interest. Although there is still a problem of representing others or speaking for others that researchers must address, the sense of crisis has diminished. The reason for this is that the ideas that objectivity is not possible, that correspondence is not an adequate theory of truth, that there is no theory-free knowledge, and so on, have become widely accepted by social researchers. Accordingly, the task now is to elaborate on and explore the implications of these realizations. It is in this sense that Norman K. Denzin and Yvonna S. Lincoln (2000) note that the crisis of representation, which marked the fourth moment of qualitative inquiry, has led to a fifth moment (the postmodern), a sixth moment (the postexperimental inquiry), and a now developing seventh moment focusing on inquiry as moral discourse.

—John K. Smith

REFERENCES

- Denzin, N., & Lincoln, Y. (2000). Introduction: The discipline and practice of qualitative research. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 1–28). Thousand Oaks, CA: Sage.
- Marcus, G., & Fischer, M. (1986). *Anthropology as cultural critique*. Chicago: University of Chicago Press.
- Smith, J. K. (1989). *The nature of social and educational inquiry: Empiricism versus interpretation*. Norwood, NJ: Ablex.

REPRESENTATIVE SAMPLE

The purpose of a sample design is to select a subset of the elements or members of a POPULATION in a way that STATISTICS derived from the SAMPLE will represent the values that are present in the population. In order to have a representative sample, a PROBABILITY method of selection must be used, and every element in the population has to have a known, nonzero chance of selection. When such a method is employed well, then the researchers can be confident that the statistics derived from the sample apply to the population within a stated range and with a stated degree of confidence.

The relevant population is theoretically defined on the basis of a specific research HYPOTHESIS. It consists of a set of elements that comprises the entire population, such as "adults residing in the United States." The easiest way to draw a probability sample is to have a list of all of the elements in the population, called a SAMPLING FRAME, and then apply a probability design to the sampling frame to produce a representative sample.

For many sampling problems, a list of all of the elements in the population exists, or at least a very good approximation of it does. In order to estimate the proportion of drivers who had an accident in the past year, a researcher might want to interview drivers and ask them whether they had an accident in the preceding 12 months. In order to produce a sample of drivers, for example, a researcher could start with a list of all individuals with a driver's license. The list might actually include all of those with licenses as of a recent date, and the list would have to be updated to include those who had received licenses since that date.

In another instance, a researcher might be interested in estimating how many American adults have a credit card. There is no existing list of all adults in the United States, so it has to be developed or approximated in the course of the study itself. This is a typical problem for those who conduct TELEPHONE SURVEYS. In this case, it is common to use a MULTISTAGE SAMPLING process that begins with a sample of telephone numbers. This is commonly generated by a computer to account

for unlisted numbers and newly assigned ones. These numbers are called, and when a residence is contacted, all of the adults residing there are listed. Through this multistage process, a CLUSTER SAMPLING of adults residing in telephone households is produced; and in the final step, one is selected at random as the respondent. That person is asked whether he or she has a credit card.

The key question for the researcher is how well the sample estimate (proportion of drivers who had a traffic accident in the past year or proportion of adults who carry a credit card) represents the population value. This is a function of two properties of the sample: its CONFIDENCE INTERVAL and confidence level. The primary advantage of using a probability method is that it allows researchers to take advantage of the NORMAL DISTRIBUTION that results from drawing repeated samples with a given characteristic from the same population. The properties of this SAMPLING DISTRIBUTION include information about the relationship between the STANDARD DEVIATION of the distribution and the size of the samples, as well as the distribution of the resulting sample estimates.

Given these properties, if we estimated from interviews with a simple random sample of 1,000 drivers that 14% of them had an accident in the past year, we would be 95% confident that the “true value” in the total population of all drivers ranged between 11% and 17% (the 14% estimate $\pm$ the three-percentage-point margin of error). The application of the probability design to a good frame ensures that the result is a representative sample with known parameters that indicate its relative precision.

—Michael W. Traugott

REFERENCES

- Kish, L. (1965). *Survey sampling*. New York: Wiley.
 Levy, P. S. (1999). *Sampling of populations: Methods and applications*. New York: Wiley.

RESEARCH DESIGN

The logical structure of the data in a research project constitutes the project's research design. Its purpose is to yield empirical data that enable relatively unambiguous conclusions to be drawn from the data.

The purpose of a research design is to test and eliminate alternative explanations of results. For example,

research to test the proposition that marriage harms the mental health of women and benefits that of men requires careful attention to research design. One design could simply compare the incidence of depression among married women with that among married men. A finding that depression rates are higher among married women than married men is *consistent* with the proposition but does nothing to eliminate alternative explanations of this finding. If never-married women are also more depressed than never-married men, then the higher depression levels of married women (compared to married men) can hardly be attributed to marriage. Similarly, if the depression rates among women are lower after they marry than before, then the greater depression levels of married women (compared to married men) are not due to marriage.

A good research design will collect the type of data that tests each of these alternatives. The more alternative explanations that are empirically eliminated, the stronger the final conclusion will be.

Most social research designs rely on the logic of *comparing groups* (Stouffer, 1950). These comparisons strengthen tests of any proposition, test alternative explanations, and bear directly on the confidence in the conclusions and interpretations drawn from the data. For example, in the above example, three different comparisons are mentioned:

- Married men with married women
- Never-married men with never-married women
- Women before and after they marry

A design based just on the first comparison would yield weak and ambiguous conclusions open to a number of alternative explanations. If the design also included data based on the other comparison groups, less ambiguous conclusions could be drawn.

Research designs vary not just in terms of which comparisons are made, but also in the nature of the groups that are compared and in the way in which group members are recruited.

The four main types of research designs (de Vaus, 2001) used in social research are the following:

- Experimental
- Longitudinal
- Cross-sectional
- Case study

Although some writers use the term *research design* to describe the *steps* in a research project or even the method of collecting data (e.g., interview, observation,

etc.), this is misleading (Yin, 1989). Research design is not simply a research plan or a method of collecting information—design is a *logical* rather than a *logistical* exercise. It is analogous to the task of an architect designing a building rather than a project manager establishing a project management schedule.

All research, regardless of the method of data collection and irrespective of whether it has a QUANTITATIVE or a QUALITATIVE emphasis, has a research design. This design may be implicit or explicit. Where it is implicit, the design is often weak in that useful comparison groups are not built into the research, thus weakening the conclusions that can be drawn from the data.

—David A. de Vaus

REFERENCES

- de Vaus, D. A. (2001). *Research design in social research*. London: Sage.
- Stouffer, S. A. (1950). Some observations on study design. *American Journal of Sociology*, 55, 355–361.
- Yin, R. K. (1989). *Case study research: Design and methods*. Newbury Park, CA: Sage.

RESEARCH HYPOTHESIS

Empirical research is done to learn about the world around us. It flows from general hunches and theories about behavior—social, biological, physical, and so on. Most research is concerned about possible relationships between variables, and hunches and theories get translated into specific statements about variables. Such statements become research hypotheses, stating how the researcher thinks the variables are related. Because it is not known whether these statements are true or not, they can be considered questions about the world. Hopefully, collecting and analyzing empirical data will provide answers.

“Do Republicans make more money than Democrats?” might be one such question as a part of a larger theory of political behavior, and a possible research hypothesis is, “Republicans make more money than Democrats.” Data can be obtained in a SAMPLE survey about how people voted and on their annual incomes.

A research hypothesis must first be translated to a statistical NULL HYPOTHESIS. The machinery of statistical inference does not make use of the research

hypothesis directly. Instead, one or more statistical parameter are identified, and possible values of the PARAMETERS are stated in such a way that they replace the research hypothesis. This produces a statistical NULL HYPOTHESIS, which then can be analyzed using methods of statistical inference, using classical statistical HYPOTHESIS TESTING. A BAYESIAN approach would avoid some of the difficulties associated with the frequentist hypothesis testing.

The translation from the research hypothesis to a more specific statistical null hypothesis can be difficult. For our example, one possible null hypothesis is that the mean population income for Republicans is equal to the mean population income for Democrats ($H_0 : \mu_r = \mu_d$). That way, the original question is shifted to a question about means, which may or may not be what the researcher has in mind.

Also, the statistical null hypothesis states that the means are equal for the two groups, whereas the research hypothesis states that one group is higher, in some way, than the other group. The focus here is on the term *equal*, even though the research hypothesis is framed in terms of *larger*. The reason for this is that the aim of most research is to prove the null hypothesis of no difference or no relationship to be wrong. If we can eliminate the possibilities that two means are equal, then we have shown that they must be different. This becomes a double-negative way of proving a research hypothesis to be true.

Even if the data show a research hypothesis to be true, the data deal only with statistical associations or differences and not CAUSAL patterns in the variables. Causality is hard to establish, and most often it is limited to data obtained from experiments. Observational data do not lend themselves easily to determining causality. One of the very few cases of causality established from observational data is the one stated on every pack of cigarettes: Smoking is bad for your health.

—Gudmund R. Iversen

REFERENCES

- Bickman, L., & Rog, D. J. (1998). *Handbook of applied social research methods*. Thousand Oaks, CA: Sage.
- Crano, W. D., & Brewer, M. B. (2002). *Principles and methods of social research* (2nd ed.). Mahway, NJ: Lawrence Erlbaum.
- Henkel, R. E. (1976). *Tests of significance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–004). Beverly Hills, CA: Sage.

RESEARCH MANAGEMENT

There are many aspects to the successful management of a research study or a research organization. Essentially, the primary goal of a research manager is to utilize available resources to best maximize the quality (i.e., RELIABILITY and VALIDITY) of the data gathered for a study or by an organization. This is neither an easy nor a straightforward process, and those who hold positions in research management will benefit from being experienced researchers in their own right. Ultimately, researcher managers must be able to aid wise decision making about the many design and operational trade-offs that are faced in planning and implementing any research study. That is, there are only finite resources available for any study, and these must be deployed wisely (from a cost-benefit standpoint).

In addition to having experience and expertise in the methodological and statistical aspects of the research that will be conducted, a successful manager or management team is likely to need knowledge and skills in fund-raising, finance, personnel management, communication and public relations, and research ethics. Without adequate financial and personnel resources, a research study will not be able to gather DATA of a level of quality that is adequate for the goal(s) of the research. Therefore, skills in fund-raising are needed to attract the resources necessary to collect valid data and to adequately analyze, interpret, and disseminate the findings. A manager also is responsible for seeing that competent staff are recruited, hired, motivated, evaluated, and retained in order to sustain the necessary work. Communication skills are needed so that the new knowledge generated by the research can be adequately disseminated. (In many instances, dissemination helps the cause of future fund-raising.) Management also must have a careful eye to all the ethical aspects of the research, especially as they relate to protecting the well-being of any humans, other animals, and/or the environment that the research may affect and that an institution's Institutional Review Board processes must sanction.

—Paul J. Lavrakas

RESEARCH QUESTION

Research questions state a research problem in a form that can be investigated and define its nature and

scope. It is through the use of such questions that choices about the focus and direction of research can be made, its boundaries can be clearly delimited, manageability can be achieved, and a successful outcome can be anticipated. One or more research questions provide the foundation on which a research project is built. Their selection and wording determines what, and to some extent, how, the problem will be studied.

There are three main types of research questions: what, why, and how. “What” questions require a descriptive answer; they are directed toward discovering and describing the characteristics of and regularities in some social phenomenon (e.g., categories of individuals, social groups of all sizes, and social processes). “Why” questions ask for either the causes of, or the reasons for, the existence of characteristics or regularities. They are concerned with understanding or explaining the relationships between events, or within social activities and social processes. “How” questions are concerned with bringing about change, and with practical outcomes and intervention (Blaikie, 2000, pp. 60–62).

These three types of research questions form a sequence: “What” questions normally precede “why” questions, and “why” questions normally precede “how” questions. We need to know what is going on before we can explain it, and we need to know why something behaves the way it does before we can be confident about introducing an intervention to change it. However, most research projects will include only one or two types of research questions—most commonly, “what” and “why” questions—and some research may deal only with “what” questions (Blaikie, 2000, pp. 61–62).

Some writers have proposed more than three types of research questions, such as “who,” “what,” “where,” “how many,” “how much,” “how,” and “why” (e.g., see Yin, 1989). However, all but genuine “why” and “how” can be transposed into “what” questions: “what individuals” in “what places,” at “what time,” in “what numbers or quantities,” and in “what ways.”

Research projects differ in the extent to which it is possible to produce precise research questions at the outset. It may be necessary to first undertake some exploratory research. In addition, what is discovered in the process of undertaking the research is likely to require a review of the research questions from time to time.

Formulating research questions is the most critical and, perhaps, the most difficult part of any RESEARCH DESIGN. However, few textbooks on research methods

give much attention to them, and some ignore this vital part of the research process entirely; those that do discuss them tend to focus on qualitative research (e.g., see Flick, 1998, and Mason, 2002).

Conventional wisdom suggests that research should be guided by one or more hypotheses. Advocates of this position suggest that research starts with the formulation of the problem, and that this is followed by the specification of one or more hypotheses, the measurement of concepts, and the analysis of the resulting variables to test the hypotheses. However, this procedure is relevant only in quantitative research that uses DEDUCTIVE logic. Although there is an important role for hypotheses in this kind of research, they do not provide the foundation for a research design, nor are they very useful for defining its focus and direction.

In some kinds of research, it is impossible or unnecessary to set out with hypotheses. What is required is the establishment of one or more research questions. The role of hypotheses is to provide tentative answers to “why” and, sometimes, “how” research questions. They are our best guesses at the answers. However, they are not appropriate for “what” questions. There is little point in hazarding guesses at a possible state of affairs. Research will produce an answer to a “what” question in due course, and no amount of guessing about what will be found is of any assistance (Blaikie, 2000, pp. 69–70).

—Norman Blaikie

REFERENCES

- Blaikie, N. (2000). *Designing social research: The logic of anticipation*. Cambridge, UK: Polity.
- Flick, U. (1998). *An introduction to qualitative research*. London: Sage.
- Mason, J. (2002). *Qualitative researching* (2nd ed.). London: Sage.
- Yin, R. K. (1989). *Case study research: Design and method* (rev. ed.). Newbury Park, CA: Sage.

RESIDUAL

Residuals measure the distance between the actual and predicted values for any estimator. If $\hat{y}$ (read y hat) is the estimated value of y , then the residual ($\hat{u}$) can be calculated by simply subtracting the true value of y from the estimated value: $\hat{u} = \hat{y} - y$. Residuals are used in a variety of ways. They are used to calculate REGRESSION lines, assess GOODNESS-OF-FIT MEASURES, and as a diagnostic in data analysis.

ORDINARY LEAST SQUARES (OLS) uses squared residuals to produce regression estimations. OLS fits a regression line to the data so the square of the residuals is minimized. Thus, OLS produces a regression line with the smallest residuals possible. The normal equations carried out by OLS produce residuals with certain properties. First, the expected values of the residuals are equal to zero. This does not mean all the residuals are equal to zero, but rather, the average of all residuals equals zero. Second, the sample COVARIANCE between each INDEPENDENT VARIABLE and the OLS residual is zero. Additionally, the sample covariance between the OLS residual and the fitted values is zero (Wooldridge, 2003).

Residuals are also used in assessing the goodness of fit of the regression model. We assess goodness of fit by how close the actual values are to the estimated regression line; the closer the values, the better the fit. Residuals are used in computing several different measures of goodness of fit. *R-SQUARED* and *ROOT MEAN SQUARE error* (root MSE) are two different measures of goodness of fit that employ residuals in their calculations.

Residuals also can tell us a considerable amount about our DATA and MODELS. Many times in the data analysis process, we examine the residuals. This can be done by looking at them in relation to the plotted regression line. If the residual is negative, we know that the regression equation overpredicts the value of the DEPENDENT VARIABLE. If the residual is positive, the regression line underpredicts the value of the dependent variable. Examining the residuals in this manner can tell us if there are certain types of observations where our model performs poorly. Damodar N. Gujarati (2003) suggests that plotting residuals is a useful exercise. In CROSS-SECTIONAL DATA, plotting the residuals can reveal problems with model MISSPECIFICATION and HETEROSKEDASTICITY. If a model lacks an important variable, or if the wrong functional form is specified, the residuals will form distinct patterns. In TIME SERIES, examining the residuals is a useful exercise for detecting heteroskedasticity and AUTOCORRELATION.

—Heather L. Ondercin

REFERENCES

- Gujarati, D. N. (2003). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.
- Wooldridge, J. M. (2003). *Introductory econometrics: A modern approach*. Cincinnati, OH: South-Western.

RESIDUAL SUM OF SQUARES

One of the key components of ANALYSIS OF VARIANCE (ANOVA) and REGRESSION, both LINEAR and multiple, is the residual sum of squares, also called sum of squared errors or error sum of squares. The residual sum of squares is the difference between the observed and predicted values of the dependent variable, reflecting the variance in the dependent variable unexplained by the linear regression model.

In ordinary least squares regression, each residual is squared in value and sums taken of all of the observed residuals squared to create the residual sum of squares, denoted SSE. The residual sum of squares indicates the sum of squares not accounted for by the linear model. It describes the variation of the observations from the prediction line. Residual sum of squares (SSE), in mathematical expression, is $\sum (Y_i - \hat{Y}_i)^2$, with Y_i = observed values and $\hat{Y}_i$ = predicted values.

A numeric example of using residual sum of squares is provided in the following narrative. In a hypothetical study of weight loss among four people, the dependent variable is the weight loss (Y) and the independent variable is the exercise (X). Each of the four people experiences a reduction in weight, measured in pounds, over a 2-week time frame, as reflected in Table 1, column Y . Column X reflects the total observed amount of exercise, measured in hours, performed by each individual over the course of 2 weeks. The column labeled $\hat{Y}$ is the predicted value of Y , based on the prediction line. The column labeled $(Y - \hat{Y})$ indicates the individual residuals between each observed versus predicted value of Y . Finally, the column $(Y - \hat{Y})^2$ provides the squared residuals. By totaling all of the squared residuals, the residual sum of squares can be determined. In this case, that value is 677.

Residual sum of squares is useful for measuring the accuracy of regression prediction. As the prediction equation improves with predicted values closer to the observed values, the residuals of each observation tend to decrease, and subsequently, the residual sum of squares gets smaller.

Residual sum of squares can be used to determine the COEFFICIENT OF DETERMINATION, denoted r^2 for bivariate models, which indicates the proportional reduction in error based on using the linear prediction equation instead of the sample mean to predict values of the dependent variable. The equation for the coefficient of

Table 1 Example of Residual Sum of Squares

Observation	X	Y	$\hat{Y}$	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$
1	7	28	14	14	196
2	15	12	30	-18	324
3	8	5	16	-11	121
4	10	14	20	-6	36

determination is $(TSS - SSE)/TSS$, with TSS = total sum of squares and SSE = residual sum of squares. For multiple regression, the same formula is used to calculate the coefficient of multiple determination, denoted R^2 .

The residual sum of squares plays a role in determining the significance of the coefficient of determination. The pertinent test statistic is called the F RATIO, and its equation is

$$\frac{\frac{R^2}{k}}{\frac{1-R^2}{n-(k+1)}},$$

with R^2 = the coefficient of determination, k = the number of explanatory variables in the model, and n = the sample size. Furthermore, the residual sum of squares can be used to compare complete and reduced models in multiple regression. The equation for the test statistic, also the F test, is

$$\frac{\frac{SSE_r - SSE_c}{k-g}}{\frac{SSE_c}{n-(k+1)}},$$

with SSE_r = residual sum of squares for the reduced model, SSE_c = residual sum of squares for the complete model, k = the number of explanatory variables in the complete model, g = the number of explanatory variables in the reduced model, and n = the sample size. This test allows the user to determine which model is better for accurately predicting the values of the dependent variable.

—Shannon Howle Schelin

See also SUM OF SQUARED ERROR

REFERENCES

- Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Lewis-Beck, M. S. (1995). *Data analysis: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-103). Thousand Oaks, CA: Sage.

RESPONDENT

A respondent is the individual chosen to provide answers to survey questions. In many cases, a one-to-one correspondence exists between the sampling units that appear on a **SAMPLING FRAME**, the respondents, and the study **POPULATION** members or units on which the researchers want to report. In these cases, researchers can choose a sample from a list of individual study population members that can and are eligible to participate in the research project. When this is the case, the sampling process uniquely identifies the respondents. For example, specific public school principals could be identified as respondents for a survey of principals' attitudes and behaviors by sampling from a list of schools, because presumably, each school has one and only one principal. In this case, the sampling unit, the respondent, and the **UNIT OF ANALYSIS** uniquely identify each other.

There are three reasons why the respondent may be different from the sampled unit:

1. The sampling units are clusters or aggregations of the individual units of analysis.
2. The survey is administered in a manner that does not allow the respondent to answer questions directly even though the respondent is capable of providing answers.
3. Some or all of the study population is incapable of providing responses.

Each of these reasons, and some solutions for them, will be addressed below.

SAMPLED UNITS ARE AGGREGATIONS

Many surveys, such as those using **RANDOM DIGIT DIALING** telephone surveys or area **PROBABILITY SAMPLING** for personal interviews (Henry, 1990; Lavrakas, 1987), identify households or other clusters of individuals from which a specific respondent must be selected. To select the person who answers the phone or the door as the respondent would bias the sample, because some family members (often females) are more likely to answer than others. Therefore, social scientists have devised several types of objective procedures to choose an eligible household member (usually 18 or over and residing in the household) to respond to the survey. Three methods are most often employed to minimize systematic bias: the Kish

procedure, the Trodahl-Carter procedure, and the last birthday procedure (Czaja, Blair, & Sebestile, 1982; O'Rourke & Blair, 1983). For example, the National Election Studies (NES) have used the Kish procedure to select an eligible voter as the respondent from each household that is identified through the sampling procedures. Interview procedures for the NES do not allow substitution for the eligible voter initially selected.

In some cases, respondents are systematically selected based upon criteria that are important for the study, such as selecting the most knowledgeable household member or a specific member of the household. This is frequently the case when the respondent is asked not only about his or her own attitudes and behaviors but also about other individuals or the collective (e.g., household). For example, Hess (1985) indicated that

preferred respondents for the Surveys of Consumer Finances, which were concerned with the collection of financial data, were heads of spending units; only when the head was unavailable for the duration of the interviewing period was an interview to be taken with the wife of the head. (p. 53)

SURVEY PROCEDURES EXCLUDE CERTAIN RESPONDENTS

Some surveys are administered in ways that preclude the sample's study population members from responding. Interviews that are conducted only in English preclude responses from non-English speakers. In these cases, the researcher has a choice to allow a surrogate to answer—for example, to ask an English-speaking family member to translate for the study population member when possible—or to exclude non-English speakers from the interviews. Either choice can **BIAS** the responses, and the researcher must decide which approach minimizes study bias or must reconsider administering the survey in languages other than English. The limitations of surveys not only apply to language differences but also to study population members with hearing impairments or speaking difficulties.

STUDY POPULATION MEMBERS UNABLE TO RESPOND

Often, the study population members, or at least some of those chosen for the sample, are incapable

of responding to the survey. For example, studies of young children often rely on parents or teachers to provide responses for the children. In some cases, such as studies of the elderly or seriously ill, some of the sampled study population members may not be able to answer the questions. Researchers must decide whether to allow surrogates to provide or assist in providing answers, or to allow NONRESPONSE BIAS from omitting the responses of those unable to provide them for themselves.

In most surveys where listed samples are used, the sampling unit is the respondent, and additional procedures are not required to specify the respondent. However, two of the most popular forms of sampling used in the social sciences, area probability sampling and random digit dialing, require additional procedures to select a respondent because these procedures identify households, not individuals. In many cases, survey researchers have eligible study population members who will be excluded from the survey without accommodation of some type. In recent years, as the non-English-speaking proportion of English-speaking countries has increased, efforts have been expanded to increase the number of individuals who can provide survey responses. Cost and the capacity to develop procedures for accommodations, translate instruments, and administer the surveys are issues for social researchers to consider as decisions about accommodation are reached.

—Gary T. Henry

REFERENCES

- Czaja, R., Blair, J., & Sebestile, J. P. (1982). Respondent selection in a telephone survey: Comparison of three techniques. *Journal of Marketing Research*, 21, 381–385.
- Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.
- Hess, I. (1985). *Sampling for social research surveys 1947–1980*. Ann Arbor, MI: Institute for Social Research.
- Lavrakas, P. J. (1987). *Telephone survey methods*. Newbury Park, CA: Sage.
- O'Rourke, D., & Blair, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research*, 20, 428–432.

RESPONDENT VALIDATION. *See*
MEMBER VALIDATION AND CHECK

RESPONSE BIAS

With a psychological rating scale or self-report, a response bias is the tendency for a person to inject idiosyncratic, systematic biases into his or her judgments. These biases introduce subjectivity that distorts the judgments. If those BIASES occur across measures of different variables, they will distort observed relations between them (see METHOD VARIANCE). It should be kept in mind that people vary in their tendencies to exhibit response biases.

With scales that deal with aspects of the self, perhaps the most well-studied response bias is SOCIAL DESIRABILITY. This is the tendency to respond to items about self in a socially acceptable direction, regardless of the truth. Social desirable response bias will be seen only in items that are socially sensitive and do not occur universally. For example, one will tend to see this bias in ratings of psychological adjustment but not job satisfaction (Spector, 1987). In addition to a response bias, social desirability has been identified as the personality variable of need for approval (Crowne & Marlowe, 1964). Individuals who have a high need for the approval of others tend to respond in a socially desirable direction on questionnaires.

With rating scales that involve judgments about others, the most common response bias is *halo*, which is the tendency to rate all dimensions for the same individual the same regardless of true variation. Thus, a person exhibiting this bias might rate one person high on all dimensions and another low on all dimensions, even though each person might vary across dimensions. This might be due to the rater making a global judgment about each person and then assuming the person fits the global judgment on all dimensions. This sort of response bias is distinct from RESPONSE SETS (e.g., leniency or severity) in that the latter are independent of the item or rating target. For example, with instructor grading of students, a response set would be the tendency to rate all students the same (e.g., to be lenient and give all high grades). A response bias would be the tendency to inject subjectivity in the rating of each individual, giving those liked inflated grades and those disliked deflated grades.

Rating bias is inferred from patterns of responses by participants, but it is not always possible to accurately distinguish bias from accurate judgments. For example, with halo bias, it has been recognized that

in many situations, there can be real relations among different dimensions. In the employee job performance domain, researchers have distinguished true halo, or real relations among dimensions of performance, from illusory halo, which involves rater biases (see Latham, Skarlicki, Irvine, & Siegel, 1993).

—Paul E. Spector

REFERENCES

- Crowne, D., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: Wiley.
- Latham, G. P., Skarlicki, D., Irvine, D., & Siegel, J. P. (1993). The increasing importance of performance appraisals to employee effectiveness in organizational settings in North America. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology 1993* (Vol. 8, pp. 87–132). Chichester, UK: Wiley.
- Spector, P. E. (1987). Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem? *Journal of Applied Psychology*, 72, 438–443.

RESPONSE EFFECTS

Response effects is a general term that covers a range of responses to SURVEY questions that is independent of the purposes of the question asked. Thus, things like the context effect of a question (i.e., what other types of related questions are in the survey?); ORDER EFFECTS (i.e., is the question before or after related questions?); memory effects (i.e., how well can the respondent remember from question to question?); wording effects (i.e., do changes in the wording of the question elicit different responses?); SOCIAL DESIRABILITY effects (i.e., are respondents giving what they think is the “right” answer?); and RESPONSE SETS (i.e., favoring a certain response category regardless of the question asked) are all examples of response effects. In a sense, these response effects are all sources of error.

—Michael S. Lewis-Beck

RESPONSE SET

With a psychological self-report or rating scale, a response set is the tendency to exhibit a particular pattern of response independent of the question being asked or the stimulus (person) being judged. For example, with a multiple-choice test format, a response

set might be to favor one of the lettered choices, for example, choice “3” or “c.”

There are several identified response sets. With self-report scales, *acquiescence* is the tendency to say yes (with yes/no items) or agree (with agree/disagree items) regardless of item content. A person exhibiting this set in the extreme will agree with both favorable and unfavorable items, although this means his or her responses are in conflict. It is possible to have a non-acquiescence set, in which the person disagrees with all items regardless of content.

There are three response sets that have been noted with scales in which people are asked to rate several dimensions of others’ behavior, such as evaluating the annual performance of employees or the abilities of students applying for graduate school. *Leniency* is the tendency to rate all individuals high on the dimension of interest. It can be seen where a rater, for example, tends to use only the top two choices on a five-point rating scale, such as instructors who award only grades of “A” or “B.” *Severity* is the opposite, where the rater gives only low ratings, such as instructors who give only “D” and “F” grades. *Central tendency* is the tendency to use only the middle of the scale, such as an instructor who tends to give mostly “C” grades and few, if any, “A” or “F” grades.

Although in theory, response sets are patterns of response independent of the item being asked or the real level of what is being judged, it seems unlikely that such sets are truly independent of content. Rohrer (1965), for example, was able to demonstrate that the acquiescence response set was not really independent of content but occurred within items of the same instrument and not across different instruments. In other words, an individual who exhibited a response set for one instrument did not exhibit it for other instruments. Spector (1987) showed that the tendency to agree with both favorable and unfavorable items was quite rare. It seems more likely that people exhibit RESPONSE BIAS, in which their patterns of response are affected by the items asked and the targets of what is being judged.

—Paul E. Spector

REFERENCES

- Rohrer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, 63, 129–156.
- Spector, P. E. (1987). Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem? *Journal of Applied Psychology*, 72, 438–443.

RETRODUCTION

This logic of inquiry is found in the version of REALISM associated with the work of Harré (1961) and Bhaskar (1979), versions also referred to as transcendental or CRITICAL REALISM. Retroduction has been identified by Harré and Bhaskar as overcoming the deficiencies of the logics of INDUCTION and deduction to offer causal explanations. They have also argued that it has been the basis for the development of past significant scientific explanations.

Retroduction entails the idea of going back from, below, or behind observed patterns or regularities to discover what produces them. Both Bhaskar and Harré rejected POSITIVISM's pattern model of explanation, that is, that explanation can be achieved by establishing regularities, or constant conjunctions. For them, establishing such regularities is only the beginning of the process. What is then required is to locate the structures or mechanisms that have produced the regularity. These structures and mechanisms are the tendencies or powers of things to act in a particular way. The capacity of a thing to exercise its powers, or the likelihood that it will, depends on whether or not the circumstances are favorable.

According to Bhaskar (1979), new mechanisms belong to a domain of reality that has yet to be observed. In order to discover these mechanisms, it is necessary to build hypothetical MODELS of them such that, if they were to exist and act in the postulated way, they would account for the phenomena being examined. These models constitute hypothetical descriptions that, it is hoped, will reveal the underlying mechanisms. Hence, these mechanisms can be known only by first constructing ideas about them. Retroduction identifies this process.

Building these hypothetical models is a creative activity involving disciplined scientific imagination and the use of analogies and metaphors. Once a model has been constructed, the researcher's task is to establish whether the structure or mechanism that it postulates actually exists. This may involve testing predictions based on the assumption that it does exist, and perhaps devising new instruments to observe it. The major value of the hypothetical model is that it gives direction to research; the retroductive researcher, unlike the inductive researcher, has something to look for.

The logic of retroduction has been used in the natural sciences in the discovery of previously unknown causal mechanisms, such as atoms and viruses. It has been applied by Harré in his ETHOGENIC approach to social psychology (see also Harré & Secord, 1972), and by Pawson in a realist form of social explanation based on a combination of the views of Harré and Bhaskar (Pawson & Tilley, 1997). The notion of retroduction and its social scientific applications are reviewed by Blaikie (1993, 2000).

Because retroduction is a relatively new notion in the social sciences, it has yet to attract serious criticism. However, given its foundation on critical realism, criticisms of the latter may also be leveled at retroduction. For the most part, these criticisms are likely to come from adherents to other PARADIGMS.

—Norman Blaikie

REFERENCES

- Bhaskar, R. (1979). *The possibility of naturalism*. Brighton, UK: Harvester.
- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Blaikie, N. (2000). *Designing social research: The logic of anticipation*. Cambridge, UK: Polity.
- Harré, R. (1961). *Theories and things*. London: Sheed & Ward.
- Harré, R., & Secord, P. F. (1972). *The explanation of social behaviour*. Oxford, UK: Basil Blackwell.
- Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage.

RHETORIC

The term *rhetoric* refers both to the persuasive character of discourse and to the long-established tradition of studying oratory. Political speeches, for example, are instances of rhetoric in the first sense because they are communications designed to persuade. Also, political speeches can be studied by the discipline of rhetoric, which will aim to demonstrate how speakers use discursive devices to try to convince their hearers.

The study of rhetoric dates back to ancient times. Aristotle and Cicero, for instance, analyzed the nature of oratory and sought to provide systematic, practical guidance for aspiring public speakers. In Europe during the Middle Ages, rhetoric was one of the three basic subjects of school curriculum, alongside mathematics

and grammar. By the time of the Renaissance, rhetoric was becoming the study of written literature, with rhetoricians devoting much effort to coding various figurative uses of language. By the late 19th century, the new disciplines of linguistics, psychology, and sociology were beginning to displace rhetoric from its preeminent place in the study of language and persuasion.

Classical rhetoric was essentially premethodological, for it did not provide analysts with a set of methodological procedures in the modern sense. This was one reason why the rhetorical tradition seemed to be in terminal decline. However, thinkers such as Kenneth Burke and Chaim Perelman were able to revive the traditions of classical rhetoric in the mid-20th century. This new rhetoric has had one very important consequence for the human and social sciences—the new rhetoricians have extended rhetoric from the traditional topics of oratory and literary texts to encompass the study of all discourse, including academic and scientific writing.

In particular, the “rhetoric of inquiry” has sought to show that the sciences and the social sciences are deeply rhetorical enterprises (Nelson, Megill, & McCloskey, 1987). Even disciplines that pride themselves on being “factual” and “scientific” use rhetorical devices in their technical reports. For example, Deirdre McCloskey applied Kenneth Burke’s rhetorical theory to the discipline of economics in her important book *The Rhetoric of Economics* (originally published in 1986 under the name of Donald McCloskey). She argues that economists habitually use figures of speech such as metaphor and metonymy, and that they employ mathematical formulae as ornamental devices of persuasion. Economists use such devices to convince readers of the factual, scientific nature of their conclusions. Other researchers have examined the rhetorical conventions of psychological, sociological, mathematical, and biological writing.

The rhetoric of inquiry shows how pervasive rhetoric is in human affairs (Billig, 1996). Inevitably, the exposure of rhetoric in academic writing has critical implications for disciplines that officially claim to be factual and thereby supposedly nonrhetorical. Moreover, the rhetoric of inquiry demonstrates that sciences and social sciences do not just comprise methodology and theory. Writing practices are crucial, although these are often treated as natural by practitioners. The extent to which the practices of writing, and thus the particular rhetorics of individual disciplines, influence

the content of theory and the practice of methodology is, at present, a matter of debate.

—Michael Billig

REFERENCES

- Billig, M. (1996). *Arguing and thinking*. Cambridge, UK: Cambridge University Press.
- McCloskey, D. N. (1986). *The rhetoric of economics*. Madison: University of Wisconsin Press.
- Nelson, J. S., Megill, A., & McCloskey, D. N. (Eds.). (1987). *The rhetoric of the human sciences*. Madison: University of Wisconsin Press.

RHO

Rho is a POPULATION parameter that represents the population correlation coefficient of the PEARSON’S CORRELATION COEFFICIENT. Suppose there are N cases in a population, and each one of the cases has a value for RANDOM VARIABLE X and for random variable Y . In other words, there are N pairs of (X, Y) in a population. The population correlation coefficient between both variables X and Y (denoted as ρ_{XY}) is defined as

$$\frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

or

$$\frac{\sum z_{X_i} z_{Y_i}}{N},$$

which are the same mathematically. σ_X^2 and σ_Y^2 are the true variances of variables X and Y , which can be obtained from

$$\sigma_X^2 = \frac{\sum (X_i - \mu_X)^2}{N}$$

and

$$\sigma_Y^2 = \frac{\sum (Y_i - \mu_Y)^2}{N},$$

where μ_X and μ_Y are the population means of variable X and variable Y , respectively. $\text{Cov}(X, Y)$ is the true covariance of the population, defined as

$$\frac{\sum (X_i - \mu_X)(Y_i - \mu_Y)}{N}.$$

z_{X_i} and z_{Y_i} are Z-SCORES of X and Y for case i , defined as

$$z_{X_i} = \frac{X_i - \mu_X}{\sigma_X}$$

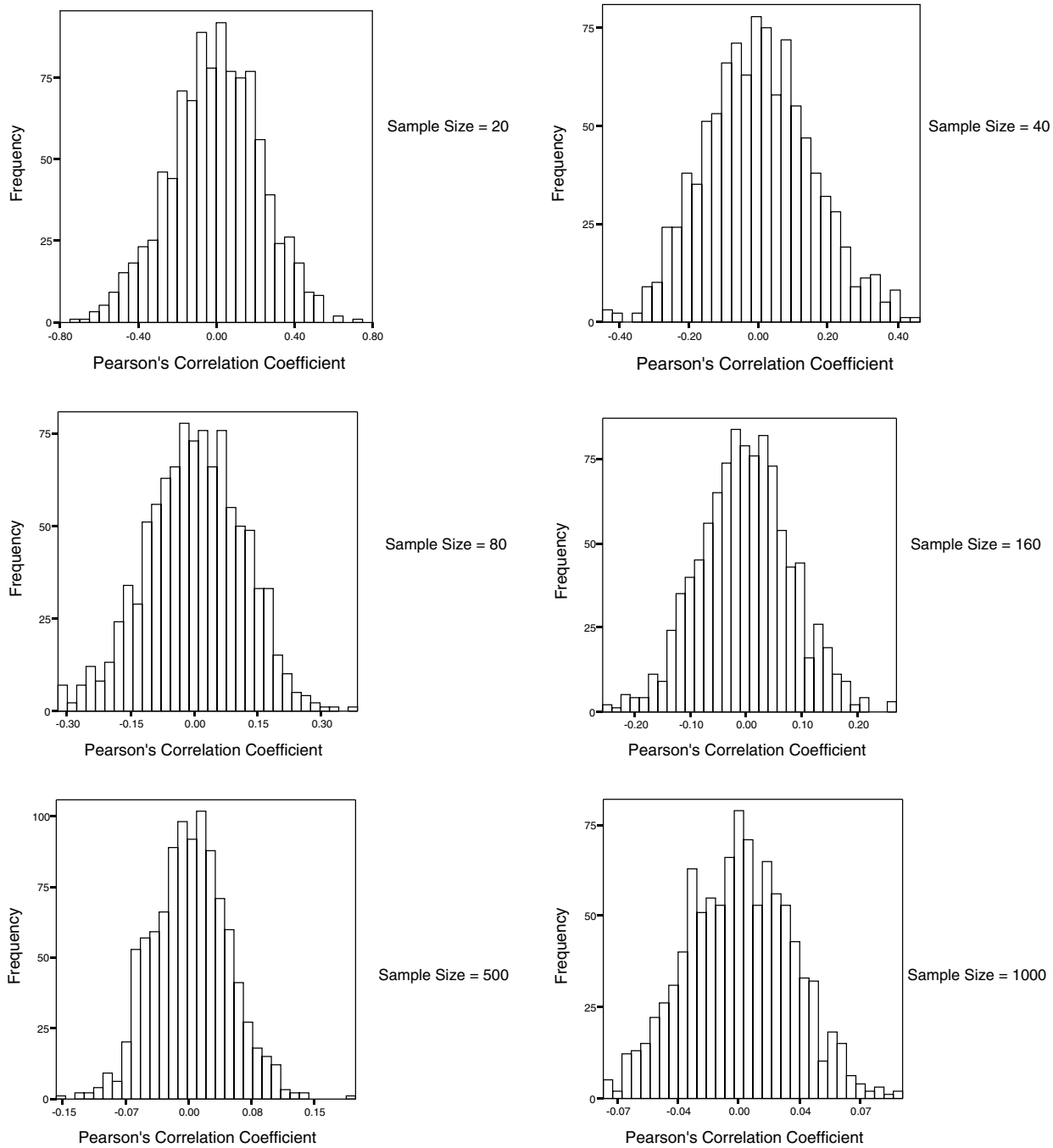


Figure 1 Frequency Distributions of Sample Correlation Coefficients Based on Six Different Sample Sizes, Given $\rho_{XY} = 0$

and

$$z_{Y_i} = \frac{Y_i - \mu_Y}{\sigma_Y}.$$

The population correlation coefficient can vary from 1 to -1 . Positive values of ρ_{XY} suggest that an increase of variable X is accompanied by an increase

of variable Y , and vice versa. In contrast, negative values of ρ_{XY} suggest that an increase of variable X is accompanied by a decrease of variable Y , and vice versa. A zero value of ρ_{XY} indicates there is no linear relationship between variables X and Y . Note that a zero value of ρ_{XY} is only a necessary rather than a

Table 1 Descriptive Results of Monte Carlo Simulations Based on Six Sample Sizes

Sample Sizes	Standard Deviation	Minimum Value	Maximum Value	Range	20th Percentile	50th Percentile	Mean	75th Percentile
20	0.2296202	-.7116160	.7270468	1.4386628	-.15944200	.004433800	-0.004665	.154351575
40	0.1538721	-.4491220	.4629634	.9120854	-.10791425	-.007544000	-0.000989	.098734075
80	0.113721	-.3216850	.3848182	.7065032	-.07750950	-.000134150	0.0002926	.076961525
160	0.0815591	-.2577240	.2692735	.5269975	-.05719025	-.001642500	-0.001815	.048942200
500	0.0451057	-.1584070	.1985341	.3569411	-.02901675	.001465750	0.0016551	.031811050
1000	0.0304901	-.0784490	.0944031	.1728521	-.02242350	.000527350	0.0000912	.021553250

sufficient condition for the INDEPENDENCE of variables X and Y . Hence, the zero value of ρ_{XY} does not imply the independence of variables X and Y . However, if both variables are determined to be independent of each other, ρ_{XY} must equal zero.

A correlation coefficient (r) based on a SAMPLE extracted from the population is often not the same as its population correlation coefficient (ρ). Imagine we randomly SAMPLE a group of 20 pairs of (X, Y) from a population, given $\rho_{XY} = 0$, and record the Pearson correlation coefficient calculated based on the formula

$$\frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sqrt{[\sum X_i^2 - \frac{(\sum X_i)^2}{n}][\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}]}}.$$

Using the same procedure, we repeatedly select a sample of 20 pairs of (X, Y) 1,000 times, followed by sampling 40, 80, 160, 500, and 1,000 pairs of (X, Y) 1,000 times, respectively. Frequency distributions of 1,000 sample correlation coefficients given six different sample sizes are depicted in Figure 1. Although the six distributions are symmetrical overall and resemble one another, the range of sample correlation coefficients increases when sample size decreases. As seen in Table 1, the range of sample correlation coefficients decreases from -0.71 to -0.08 and from 0.72 to 0.09 when the sample size increases from 20 to 1,000. Furthermore, when the sample size increases, the means of the sample correlation coefficients are closer to the population rho, given $\rho_{XY} = 0$. Comparisons among distributions when ρ_{XY} equals nonzero values will be described in detail in Pearson's correlation coefficient.

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

- Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage.
Hays, W. L. (1994). *Statistics* (5th ed.). New York: Harcourt Brace.

ROBUST

Robust in statistical estimation means *insensitive* to *small* departures from the idealized assumptions for which the ESTIMATOR is optimized. The word “small” can mean either a *small* departure for all or almost all the data points, or large departures for a proportionally *small* number of data points.

—Tim Futing Liao

ROBUST STANDARD ERRORS

Statistical inferences are based on both observations and assumptions about underlying distributions and relationships between variables. If a model is relatively insensitive to small deviations from these ASSUMPTIONS, its PARAMETER ESTIMATIONS and STANDARD ERRORS can be considered robust. In some cases, the violation of an assumption affects only the standard errors. For example, violation of the ORDINARY LEAST SQUARES (OLS) assumptions of nonconstant error variance and independence produces UNBIASED parameter estimates but inconsistent standard errors. In other words, the regular OLS standard errors are not reliable, thus calling into question tests of statistical significance. One way of overcoming this problem is to use robust standard errors.

Robust standard errors were first developed to compensate for the existence of an unknown pattern of HETEROSKEDASTICITY that cannot be totally eliminated. There are several methods for calculating heteroskedasticity-consistent standard errors; these are known variously as White, Eicker, or Huber standard errors. Most of these are variants on the method originally proposed by White (1980), which uses the heteroskedasticity-consistent covariance matrix (HCCM). The HCCM matrix estimator is

$$V(b) = (X'X)^{-1} X' \hat{\Phi} X (X'X)^{-1},$$

where $\hat{\Phi} = \text{diag}[e_i^2]$, and the e_i are the OLS residuals.

Because the HCCM is found without a formal model of the heteroskedasticity, relying instead on only the regressors and residuals from the OLS for its computation, the White correction method can be easily adapted to many different applications. For example, robust standard errors can also be used to relax the independence assumption for data from clustered samples. Moreover, robust standard errors (e.g., Newey-West standard errors) can also improve statistical inference from pooled TIME-SERIES data with AUTOCORRELATED errors and from PANEL data.

Some cautions about the use of robust standard errors are necessary. First, not all robust estimators of variance perform equally well with small sample sizes (for a detailed discussion, see Long & Ervin, 2000). Perhaps more importantly, robust standard errors should not be seen as a substitute for careful model specification. In particular, if the pattern of heteroskedasticity is known, it can often be corrected more effectively—and the model estimated more efficiently—using WEIGHTED LEAST SQUARES. Similarly, although robust standard errors can be used for clustered data, greater information about the pattern in such data—such as the variance due to the clustering—can be obtained using MULTILEVEL modeling.

Finally, REGRESSION with robust standard errors should not be confused with “robust regression,” which refers to regression models that are resistant to skewed distributions and influential cases. With such models, both the estimates and the standard errors are different from those obtained from OLS regression.

—Robert Andersen

REFERENCES

Huber, P. (1981). *Robust statistics*. New York: Wiley.

Long, J. S., & Ervin, L. H. (2000). Using heteroskedasticity consistent standard errors in the linear regression model. *American Statistician*, 54, 217–224.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55, 703–708.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817–838.

ROLE PLAYING

Role playing is a technique commonly used by researchers studying interpersonal behavior in which researchers assign research participants to particular roles and instruct those participants to act *as if* a set of conditions were true.

The use of role taking has a long history in social science research and was used in some of the early classic social psychological experiments by Kurt Lewin (1939/1997), Stanley Milgram (1963), and Phillip Zimbardo (Haney, Banks, & Zimbardo, 1973). For example, Zimbardo and colleagues set up a mock prison in the basement of the Stanford psychology department and randomly assigned undergraduate volunteers to serve in the role of either prisoner or guard. To increase the realism of the role assignments, those assigned to be prisoners were “arrested” in their dorms or classes, required to wear prisoners’ uniforms, and held in makeshift “cells” in the mock prison. Those assigned to be guards wore their street clothes and were assigned props such as mirrored sunglasses, whistles, and batons. The responses to this role assignment were dramatic. The behavior of those assigned to be guards became more authoritarian and more aggressive, whereas the behavior of those assigned to be prisoners became more passive and more dependent. Ultimately, the researchers halted the experiment prematurely when several of the mock prisoners began to experience physiological symptoms of severe stress. In a more contemporary example, Johnson (1994) made use of a role-playing experiment to show that individuals assigned to be supervisors of a mock organization exhibited different conversational behaviors from individuals assigned to be subordinates in that organization.

The process of role playing in a social experiment can be likened to a COMPUTER SIMULATION, where the human mind serves as the processor rather than a

computer. These techniques capitalize on the human ability to engage in creative role taking—to implement certain rules of thought, feeling, and behavior—under certain assumed conditions. That humans might be effective at playing arbitrarily assigned roles is bolstered by Goffman's (1959) ideas about all social behavior being a form of dramatic performance. The use of role playing in research presumes that individuals have the capacity to effectively imagine how they might behave in a given set of circumstances. It also presumes that there is a relationship between how individuals feel and act under these simulated circumstances and how they might feel and act under the actual circumstances they are simulating. However, the VALIDITY of role playing in social science research does not depend on individuals' abilities to behave precisely the way that they might in the actual circumstances being simulated. Rather, the strength of most research that makes use of role assignments is in the comparison of behaviors between individuals who have been RANDOMLY ASSIGNED to occupy different roles. Thus, it is more often the observed *differences* between the behaviors, thoughts, and feelings of those occupying the different roles that is of more interest than the actual character of the role-based behavior.

—Dawn T. Robinson

REFERENCES

- Goffman, E. (1959). *Presentation of self in everyday life*. Garden City, NY: Doubleday.
- Haney, C., Banks, C., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology*, 1, 69–97.
- Johnson, C. (1994). Gender, legitimate authority, and leader-subordinate conversations. *American Sociological Review*, 59, 122–135.
- Lewin, K. (1997). Experiments in social space. In G. W. Lewin (Ed.), *Resolving social conflicts: Field theory in social science* (pp. 59–67). Washington, DC: American Psychological Association. (Originally published in 1939)
- Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.

ROOT MEAN SQUARE

Root mean square error (RMSE) is a measure of how well an estimator performs. If $\hat{\beta}$ is an estimator of β , the RMSE of $\hat{\beta}$ is given by

$$\text{RMSE}(\hat{\beta}) = E[(\hat{\beta} - \beta)^2]. \quad (1)$$

We can immediately note that if $\beta = E[\hat{\beta}]$, then this is identical to the variance of $\hat{\beta}$. Rewriting this and simplifying the right-hand side provides the following decomposition:

$$\begin{aligned} \text{RMSE}(\hat{\beta}) &= E[(\hat{\beta} - E[\hat{\beta}] + E[\hat{\beta}] - \beta)^2] \\ &= E[(\hat{\beta} - E[\hat{\beta}])^2 + (E[\hat{\beta}] - \beta)^2 \\ &\quad - 2(\hat{\beta} - E[\hat{\beta}])(E[\hat{\beta}] - \beta)] \\ &= E[(\hat{\beta} - E[\hat{\beta}])^2] + E[(E[\hat{\beta}] - \beta)^2] \\ &\quad - 2E[(\hat{\beta} - E[\hat{\beta}])(E[\hat{\beta}] - \beta)]. \end{aligned} \quad (2)$$

The last expression reduces to 0, leaving:

$$\text{RMSE}(\hat{\beta}) = \text{Var}(\hat{\beta}) + [\text{bias}(\hat{\beta})]^2. \quad (3)$$

Again, note that if $\hat{\beta}$ is an unbiased estimator of β , then $\text{RMSE}(\hat{\beta}) = \text{Var}(\hat{\beta})$.

This leads to a fundamental comparison between two estimators: $\hat{\beta}$ and $\hat{\beta}^*$ in cases where $\hat{\beta}$ is an UNBIASED estimator of β and $\hat{\beta}^*$ is a BIASED estimator of β , but has lower variance than $\hat{\beta}$. In this case, depending upon the relative VARIANCES of the two estimators, and the size of the bias of $\hat{\beta}^*$, $\hat{\beta}^*$ could have a lower RMSE than $\hat{\beta}$. In this case, we might prefer the biased estimator to the unbiased estimator if we want to minimize RMSE!

RMSE AND OLS

RMSE has another common usage. Some statistics packages will print out a value for RMSE with the output of OLS regressions, and other linear models. It is what is more commonly known as the standard error of the regression:

$$\text{RMSE} = \sqrt{\frac{\sum (\hat{Y} - Y)^2}{n - k}}. \quad (4)$$

This is an estimator of the STANDARD DEVIATION of the POPULATION disturbance. Or, if we have a typical OLS setup, where

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon, \quad (5)$$

then RMSE is an estimate of σ , where σ^2 is the variance of ε .

Thus, RMSE provides information about one component of the errors we make with estimates: It is a measure of the STOCHASTIC component of error. RMSE is measured in the units of the dependent variable.

It gives us a measure of how well our model would fit the data-generating process if we knew the true PARAMETERS of the model. A large value of RMSE implies that a large amount of variance in the dependent variable is not explained by the model. And it means that even if we obtained very precise estimates of the model parameters, any PREDICTION we made would still have a large range of uncertainty around it.

RMSE is one component of the error of any point prediction made using MODEL estimates. The UNCERTAINTY about such a point prediction also needs to take into account the estimation error about the parameter estimates being used. Thus, in computing the uncertainty about any point estimate, $\hat{Y}_i$, the uncertainty in our estimates of $\hat{\beta}$ that go into Y_i could dwarf the stochastic uncertainty contained in RMSE.

RMSE is thus a measure of GOODNESS OF FIT. But it is a measure of goodness of fit of the model, not of the estimates. It is often compared to an alternative measure of fit, R -SQUARED (R^2). However, R^2 contains different information from RMSE. Whereas RMSE is an estimator of the stochastic component of variance, R^2 is a measure of the fit between the predicted and the observed values of Y in the sample. Mathematically, the two are related as follows:

$$\text{RMSE}^2 = \text{Var}(Y) - \text{Var}(\hat{Y}), \quad (6)$$

$$R^2 = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)}. \quad (7)$$

So, if we want to know how well we are able to fit the given sample of data, R^2 can be a useful measure. But to understand the amount of stochastic variance present, we need to look at RMSE. R^2 will vary depending upon the variance in the sample Y s, whereas RMSE will remain a consistent estimator of the stochastic variance across different samples. The distinction between RMSE and R^2 is described in more detail in Achen (1977) and King (1990).

—Jonathan Nagler

REFERENCES

- Achen, C. H. (1977). Measuring representation: Perils of the correlation coefficient. *American Journal of Political Science*, 21, 805–815.
- King, G. (1990). Stochastic variation: A comment on Lewis-Beck and Skalaban's "The R-Squared." *Political Analysis*, 2, 185–200.

ROTATED FACTOR. See FACTOR ANALYSIS

ROTATIONS

The different methods of FACTOR ANALYSIS first extract a set of factors from a data set. These factors are almost always orthogonal and are ordered according to the proportion of the VARIANCE of the original data that these factors explain. In general, only a (small) subset of factors is kept for further consideration, and the remaining factors are considered either irrelevant or nonexistent (i.e., they are assumed to reflect measurement error or noise).

In order to make the interpretation of the factors that are considered relevant, the first selection step is generally followed by a rotation of the factors that were retained. Two main types of rotation are used: orthogonal, when the new axes are also orthogonal to each other, and oblique, when the new axes are not required to be orthogonal to each other. Because the rotations are always performed in a subspace (the so-called factor space), the new axes will always explain *less* variance than the original factors (which are computed to be optimal), but obviously, the part of variance explained by the total subspace after rotation is the same as it was before rotation (only the partition of the variance has changed). Because the rotated axes are not defined according to a statistical criterion, their *raison d'être* is to facilitate the interpretation.

In this entry, the rotation procedures are illustrated using the loadings of variables analyzed with PRINCIPAL COMPONENTS ANALYSIS (the so-called R -mode), but the methods described here are valid also for other types of analysis and when analyzing the subjects' scores (the so-called Q -mode).

Before proceeding further, it is important to stress that because the rotations always take place in a subspace (i.e., the space of the retained factors), the choice of this subspace strongly influences the result of the rotation. Therefore, in the practice of rotation in factor analysis, the strong recommendation is to try several sizes for the subspace of the retained factors in order to assess the robustness of the interpretation of the rotation.

NOTATIONS

To make the discussion more concrete, rotations within the principal components analysis (PCA) framework are described. PCA starts with a data matrix denoted $\mathbf{Y}$ with I rows and J columns, where each row represents a unit (in general subjects) described by J measurements that are almost always expressed as Z -scores. The data matrix is then decomposed into scores (for the subjects) or components and loadings for the variables (the loadings are, in general, correlations between the original variables and the components extracted by the analysis). Formally, PCA is equivalent to the singular value decomposition of the data matrix as

$$\mathbf{Y} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T, \quad (1)$$

with the constraints that $\mathbf{Q}^T\mathbf{Q} = \mathbf{P}^T\mathbf{P} = \mathbf{I}$ (where $\mathbf{I}$ is the identity matrix), and $\mathbf{\Delta}$ is a diagonal matrix with the so-called singular values on the diagonal. The orthogonal matrix $\mathbf{Q}$ is the matrix of loadings (or projections of the original variables onto the components), where one row stands for one of the original variables and one column for one of the new factors. In general, the subjects' score matrix is obtained as $\mathbf{F} = \mathbf{P}\mathbf{\Delta}$ (some authors use $\mathbf{P}$ as the scores).

ROTATING: WHEN AND WHY?

Most of the rationale for rotating factors comes from Thurstone (1947) and Cattell (1978), who defended its use because this procedure simplifies the factor structure and therefore makes its interpretation easier and more reliable (i.e., easier to replicate with different data samples).

Thurstone suggested five criteria to identify a simple structure. According to these criteria, still often reported in the literature, a matrix of loadings (where the rows correspond to the original variables and the columns to the factors) is simple if

1. each row contains at least one zero
2. for each column, there are at least as many zeros as there are columns (i.e., number of factors kept)
3. for any pair of factors, there are some variables with zero loadings on one factor and large loadings on the other factor
4. for any pair of factors, there is a sizable proportion of zero loadings
5. for any pair of factors, there is only a small number of large loadings.

Rotations of factors can (and used to) be done graphically, but are mostly obtained analytically and necessitate the mathematical specification of the notion of simple structure in order to implement it with a computer program.

ORTHOGONAL ROTATION

An orthogonal rotation is specified by a rotation matrix denoted $\mathbf{R}$, where the rows stand for the original factors and the columns for the new (rotated) factors. At the intersection of row m and column n we have the cosine of the angle between the original axis and the new one: $r_{m,n} = \cos \theta_{m,n}$. For example, the rotation illustrated in Figure 1 will be characterized by the following matrix:

$$\begin{aligned} \mathbf{R} &= \begin{bmatrix} \cos \theta_{1,1} & \cos \theta_{1,2} \\ \cos \theta_{2,1} & \cos \theta_{2,2} \end{bmatrix} \\ &= \begin{bmatrix} \cos \theta_{1,1} & -\sin \theta_{1,1} \\ \sin \theta_{1,1} & \cos \theta_{1,1} \end{bmatrix}, \end{aligned} \quad (2)$$

with a value of $\theta_{1,1} = 15$ degrees. A rotation matrix has the important property of being orthonormal because it corresponds to a matrix of direction cosines and therefore $\mathbf{R}^T\mathbf{R} = \mathbf{I}$.

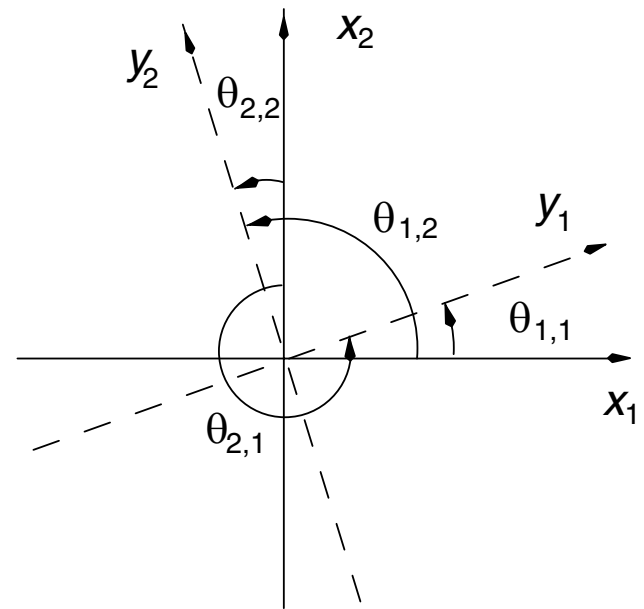


Figure 1 An Orthogonal Rotation in Two Dimensions; the Angle of Rotation Between an Old Axis m and a New Axis n Is Denoted By $\theta_{m,n}$

VARIMAX

VARIMAX, which was developed by Kaiser (1958), is indubitably the most popular rotation method by far. For varimax, a simple solution means that each factor has a small number of large loadings and a large number of zero (or small) loadings. This simplifies the interpretation because, after a varimax rotation, each original variable tends to be associated with one (or a small number of) factors, and each factor represents only a small number of variables. In addition, the factors often can be interpreted from the opposition of few variables with positive loadings to few variables with negative loadings.

Formally, varimax searches for a rotation (i.e., a linear combination) of the original factors such that the *variance* of the loadings is maximized, which amounts to maximizing

$$V = \sum (q_{j,\ell}^2 - \bar{q}_{j,\ell}^2)^2, \quad (3)$$

with $q_{j,\ell}^2$ being the squared loading of the j th variable on the ℓ factor, and $\bar{q}_{j,\ell}^2$ being the mean of the squared loadings. (For computational stability, each of the loadings matrices is generally scaled to length one prior to the minimization procedure.)

OTHER ORTHOGONAL ROTATIONS

There are several other methods for orthogonal rotation such as the quartimax rotation, which minimizes the number of factors needed to explain each variable, and the equimax rotation, which is a compromise between varimax and quartimax. Other methods exist, but none approaches varimax in popularity.

AN EXAMPLE OF ORTHOGONAL VARIMAX ROTATION

To illustrate the procedure for a varimax rotation, suppose that we have five wines described by the average rating of a set of experts on their hedonic dimension—how much the wine goes with dessert and how much the wine goes with meat; each wine is also described by its price, its sugar and alcohol content, and its acidity. The data are given in Table 1.

A PCA of this table extracts four factors (with eigenvalues of 4.7627, 1.8101, 0.3527, and 0.0744, respectively), and a two-factor solution (corresponding to the components with an eigenvalue larger than unity)

Table 1 An (Artificial) Example for PCA and Rotation; Five Wines Are Described by Seven Variables

		<i>For</i>	<i>For</i>				
	<i>Hedonic</i>	<i>Meat</i>	<i>Dessert</i>	<i>Price</i>	<i>Sugar</i>	<i>Alcohol</i>	<i>Acidity</i>
Wine 1	14	7	8	7	7	13	7
Wine 2	10	7	6	4	3	14	7
Wine 3	8	5	5	10	5	12	5
Wine 4	2	4	7	16	7	11	3
Wine 5	6	2	4	13	3	10	3

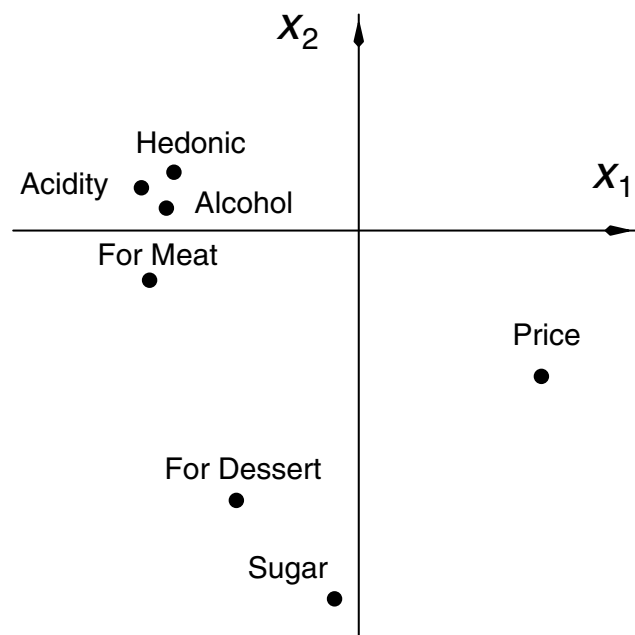


Figure 2 PCA, Original Loadings of the Seven Variables for the Wine Example

explaining that 94% of the variance is kept for rotation. The loadings are given in Table 2.

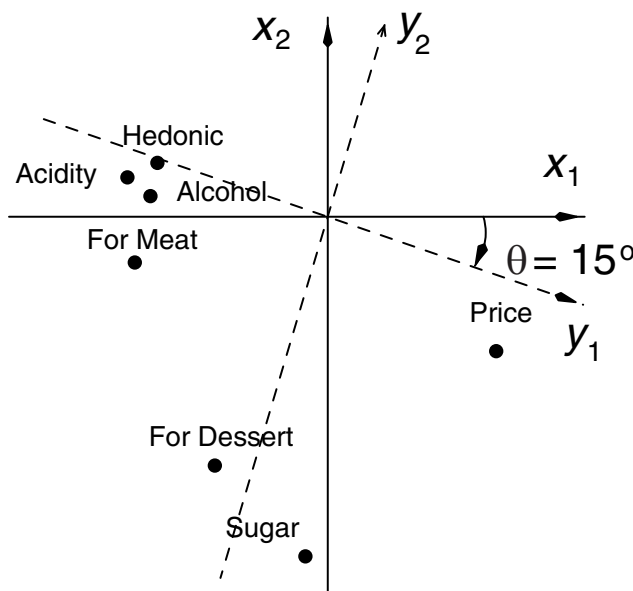
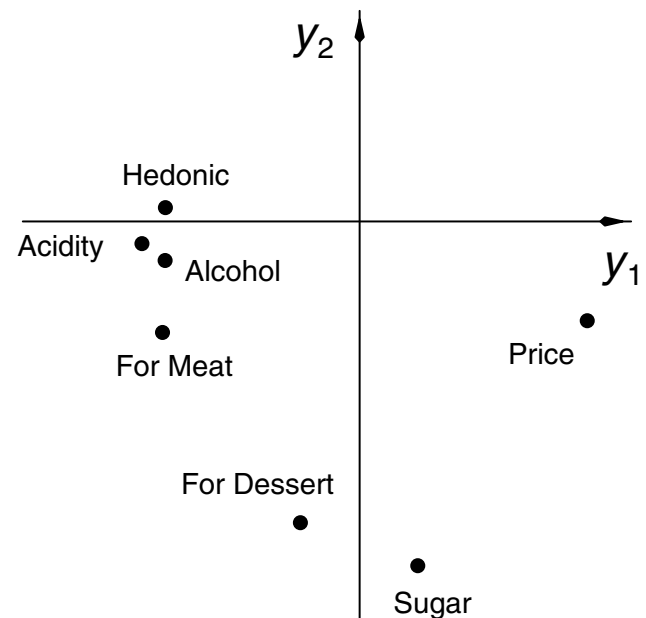
The varimax rotation procedure applied to the table of loadings gives a clockwise rotation of 15 degrees (corresponding to a cosine of .97). This gives the new set of rotated factors shown in Table 3. The rotation procedure is illustrated in Figures 2, 3, and 4. In this example, the improvement in the simplicity of the interpretation is somewhat marginal, maybe because the factorial structure of such a small data set is already very simple. The first dimension remains linked to price, and the second dimension now appears more clearly as the dimension of sweetness.

Table 2 Wine Example: Original Loadings of the Seven Variables on the First Two Components

	<i>Hedonic</i>	<i>For Meat</i>	<i>For Dessert</i>	<i>Price</i>	<i>Sugar</i>	<i>Alcohol</i>	<i>Acidity</i>
Factor 1	−0.3965	−0.4454	−0.2646	0.4160	−0.0485	−0.4385	−0.4547
Factor 2	0.1149	−0.1090	−0.5854	−0.3111	−0.7245	0.0555	0.0865

Table 3 Wine Example: Loadings, After Varimax Rotation, of the Seven Variables on the First Two Components

	<i>Hedonic</i>	<i>For Meat</i>	<i>For Dessert</i>	<i>Price</i>	<i>Sugar</i>	<i>Alcohol</i>	<i>Acidity</i>
Factor 1	−0.4125	−0.4057	−0.1147	0.4790	0.1286	−0.4389	−0.4620
Factor 2	0.0153	−0.2138	−0.6321	−0.2010	−0.7146	−0.0525	−0.0264

**Figure 3** PCA, the Loading of the Seven Variables for the Wine Example Showing the Original Axes and the New (Rotated) Axes Derived From Varimax**Figure 4** PCA, The Loadings After Varimax Rotation of the Seven Variables for the Wine Example

OBLIQUE ROTATIONS

In oblique rotations, the new axes are free to take any position in the factor space, but the degree of correlation allowed among factors is, in general, small because two highly correlated factors are better interpreted as only one factor. Therefore, oblique rotations relax the orthogonality constraint in order to gain simplicity in the interpretation. They were strongly recommended by Thurstone but are used more rarely than their orthogonal counterparts.

For oblique rotations, the promax rotation has the advantage of being fast and conceptually simple. Its name derives from procrustean rotation because it tries to fit a *target* matrix that has a simple structure. It necessitates two steps. The first step defines the target matrix, almost always obtained as the result of a varimax rotation whose entries are raised to some power (typically between 2 and 4) in order to “force” the structure of the loadings to become bipolar. The second step is obtained by computing a least squares fit from the varimax solution to the target matrix.

Table 4 Wine Example: Promax Oblique Rotation Solution; Correlation of the Seven Variables on the First Two Components

	<i>Hedonic</i>	<i>For Meat</i>	<i>For Dessert</i>	<i>Price</i>	<i>Sugar</i>	<i>Alcohol</i>	<i>Acidity</i>
Factor 1	−0.8783	−0.9433	−0.4653	0.9562	0.0271	−0.9583	−0.9988
Factor 2	−0.1559	−0.4756	−0.9393	−0.0768	−0.9507	−0.2628	−0.2360

Even though, in principle, the results of oblique rotations could be presented graphically, they are almost always interpreted by looking at the correlations between the rotated axis and the original variables. These correlations are interpreted as loadings.

For the wine example, using as target the varimax solution raised to the fourth power, a promax rotation gives the loadings shown in Table 4. From these loadings, one can find that the first dimension once again isolates the price (positively correlated to the factor) and opposes it to the other variables: alcohol, acidity, “going with meat,” and hedonic component. The second axis shows only a small correlation with the first one ($r = -.02$). It corresponds to the “sweetness/going with dessert” factor. Once again, the interpretation is quite similar to the original PCA because the small size of the data set favors a simple structure in the solution.

Oblique rotations are much less popular than their orthogonal counterparts, but this trend may change as new techniques, such as *independent component analysis*, are developed. This recent technique, originally created in the domain of signal processing and NEURAL NETWORKS, derives directly from the data matrix an oblique solution that maximizes statistical independence.

—Hervé Abdi

REFERENCES

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.

Kim, J. O., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues*. Beverly Hills, CA: Sage.

Kline, P. (1994). *An easy guide to factor analysis*. Thousand Oaks, CA: Sage.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Reyment, R. A., & Jöreskog, K. G. (1993). *Applied factor analysis in the natural sciences*. Cambridge, UK: Cambridge University Press.

Rummel, R. J. (1970). *Applied factor analysis*. Evanston, IL: Northwestern University Press.

ROUNDING ERROR

It is customary to round estimates off, up, or down. If an estimate is rounded off, then it is left with fewer digits. For example, an estimated income of \$67,542.67 might be rounded off to \$67,543, thus losing two digits. Also, note that, here, income is rounded up; the last 67 cents (.67) is rounded up to a whole dollar. Of course, if the original estimate had been \$67,542.38, we would have rounded down, to \$67,542. In all these instances of rounding—off, up, or down—some error is introduced. In the first and second instances, the error is .67; in the third instance, it is .38. A common example of rounding error occurs when column percentages are summed in a table, for they frequently sum to 99% or 101%, instead of the correct 100%. When that occurs, the analyst usually includes a footnote indicating that the percentages do not sum to 100 because of rounding error. The presence of rounding error is routine, not to say inevitable, and rarely causes difficulty, because the error is small and random. It is, of course, possible to round too much. For example, suppose we rounded the above income number to \$70,000. Now that income estimate contains a rounding error of nearly \$2,500. Thus, rounding error has increased, but is this too much error? The answer to that question depends on the RESEARCH QUESTION asked and the degree of precision the analyst desires. At the other extreme, it is possible to not round enough. For example, in most income studies, we would not need to know income measured to the penny. Indeed, to claim that we could estimate income to the penny may be giving a false sense of precision, denying the greater error inherent in the MEASUREMENT.

—Michael S. Lewis-Beck

ROW ASSOCIATION. See
ASSOCIATION MODEL

ROW-COLUMN ASSOCIATION.

See ASSOCIATION MODEL

R-SQUARED

The R -squared (R^2) measures the explanatory or predictive power of a REGRESSION model. It is a GOODNESS-OF-FIT MEASURE, indicating how well the linear regression equation fits the data.

In regression analysis, it is important to evaluate the performance of the estimated regression equation. To what extent does it account for the phenomenon under study? The R^2 is the leading performance measure for a simple or MULTIPLE REGRESSION model. Suppose a policy analyst is studying public school expenditures in the 50 American states. The analyst posits a simple regression model, $Y = a + bX + e$, where Y is the DEPENDENT VARIABLE of per-pupil public school expenditures (in thousands of dollars) in each state in the year 2000, X is the INDEPENDENT VARIABLE of urbanization (percentage of population in cities larger than 25,000, as of the 2000 census), and e is the error term. The model argues that state public school outlays are accounted for, in part, by urbanization. ORDINARY LEAST SQUARES (OLS) estimation yields, hypothetically, the following results: $Y = 2.11 + .60X + e$, suggesting that for every additional percentage of urban population, a state expects to spend \$600 more. The R^2 , or coefficient of determination, for the equation is .42, which says that 42% of the variation in state public school expenditures is explained, or at least predicted, by urbanization.

The R^2 shows the gain in predicting Y knowing X , as opposed to not knowing X . Suppose that the analyst knows the scores on Y , but not the states to which they are attached, and tries to predict school expenditures state by state. The best guess, the one that minimizes the error, would always be the average of Y , that is, Y_m . This guess will be way off for most states, with a great distance from the observed expenditure score to the average expenditure score, that is, $(Y - Y_m)$. Adding all these distances together (after squaring to overcome the canceling out of the plus and minus signs), they represent the total prediction error not knowing X ; that is, total sum of squared deviations ($TSS = \sum(Y - Y_m)^2$).

Regression analysis promises reduction of this prediction error through knowledge of X and its linear

relationship to Y . For example, knowing a state's score on X is 30% urban, the prediction, $\hat{Y}$, is 20.21 ($\hat{Y} = 2.11 + 18.10 = 20.21$), not Y_m . The distance from the predicted Y to the mean Y , $(\hat{Y} - Y_m)$, is the improvement over the baseline prediction made possible by knowing X . This distance, squared and summed for all observations, is the reduction in prediction error attributable to application of the regression line; that is, regression sum of squared deviations ($RSS = \sum(\hat{Y} - Y_m)^2$). Unless the regression line predicts each case perfectly, some error will still remain: the distance from the observed Y and the predicted Y . These distances, squared then summed, represent the variation in Y that remains unexplained; that is, error sum of squared deviations ($ESS = \sum(Y - \hat{Y})^2$).

Total variation in the dependent variable, TSS, thus has two unique parts: RSS accounted for by the regression and ESS not accounted for by the regression. The R^2 reflects the regression portion as a share of the total, $RSS/TSS = R^2$. The statistic ranges from 1.0, when all variation is accounted for, to .00, when no variation is accounted for. With real-world data, the R^2 seldom reaches these extreme values. In the above example, we may say that 42% of the variation in public school expenditures is explained by urbanization. It is important to remember that this explanation may be more "statistical" than "causal." It could be that urbanization helps predict public school expenditure but does not really explain it in a theoretical sense. In such cases, it is more cautious, and perhaps more correct, to say that X merely "accounts for" so much variation in Y . In the bivariate regression case, the R^2 equals the CORRELATION coefficient squared (the r^2).

With a multiple regression equation, the R^2 is referred to as the coefficient of multiple determination. Again, it measures the linear goodness of fit of the model, ranging from 1.0 to .00. For example, when the multiple regression model, $Y = a + bX + cZ + e$, is estimated with OLS, the R^2 indicates the fit of a plane to the points in a three-dimensional space. For more than two independent variables, a hyperplane is fitted. The MULTIPLE CORRELATION coefficient, symbolized by R , is the square root of the R^2 and indicates the correlation of the ensemble of independent variables with the dependent variable.

When an independent variable is added to a regression model, the R^2 always increases. Some of this increase is due to chance, for adding independent variables uses DEGREES OF FREEDOM. Especially as the number of independent variables approaches the

sample size, the R^2 estimate will exaggerate the real fit of the model to the data. Most analysts report the ADJUSTED R -SQUARED for a multiple regression model, along with, or instead of, the R^2 itself. The adjusted R^2 will be smaller in magnitude than the R^2 , but seldom much smaller. Unlike the R^2 , the adjusted R^2 can show decreases as independent variables are added to a model. (Rarely, the adjusted R^2 can be negative, when the unadjusted fit is extremely low in the first place. The unadjusted R^2 has the possibility of being negative only if the constant term is forced to be zero.)

Analysts generally prefer a high R^2 , sensing that a better fit means a better explanation. However, that may or may not be true. R^2 maximization should not be an end in itself. Blind inclusion into a regression model of the variables that have a high correlation with the dependent variable may yield a high R^2 . But when that model is estimated on a subsequent sample, that high fit will likely plummet, because much of it came from chance variation in the first sample. A moderate, even low, R^2 is not necessarily bad. Certain kinds of data, such as public opinion surveys, do not generally admit of high fits. Regardless of data type, a low fit may simply mean that only a small portion of Y can be explained, the rest being random. There is a case where a low R^2 may be a bad sign, the case of NONLINEARITY. If the relationship follows a curve rather than a straight line, then linear specification and the R^2 assessment of model performance is inappropriate. Finally, a low R^2 for any model suggests the model will predict poorly.

The R^2 has been subjected to various criticisms. Some consider the baseline for prediction

comparison— Y_m —to be naive. Usually, however, it is difficult to come up with a better alternative baseline. Similarly, some object that it assesses only the linearity of model fit. However, it is often hard to improve on the linear specification in practice. A further issue is whether the R^2 of different models can be tested for STATISTICALLY SIGNIFICANT differences, in order to assess rival model specifications. This can be done if one model is possibly a nested subset of another larger model. Legitimate use of such significance tests depends partly on whether the R^2 is held to be a descriptive or inferential statistic. A few analysts reject the R^2 altogether, preferring the STANDARD ERROR OF ESTIMATE (SEE) of Y as a measure of goodness of fit. The difficulty with using the SEE measure alone is that it has no conventional theoretical boundary at the upper end (e.g., 1.0). As SEE gets larger, one knows there is more prediction error, but one does not know whether it has reached its maximum. In general, the two measures of fit should both be reported because they say somewhat different things.

—Michael S. Lewis-Beck

See also COEFFICIENT OF DETERMINATION

REFERENCES

- Lewis-Beck, M. S., & Skalaban, A. (1990). The R -squared: Some straight talk. *Political Analysis*, 2, 153–170.
- Pindyck, R. S., & Rubinfeld, D. L. (1991). *Econometric models and economic forecasts* (3rd ed.). New York: McGraw-Hill.
- Weisberg, S. (1985). *Applied linear regression*. New York: Wiley.

S

SAMPLE

A sample is a subset of the elements or members of a POPULATION. It is employed to collect information from or about the elements in a way that the resulting information can represent the population from which it was drawn. Sampling is used as an efficient and cost-effective way to collect DATA that can both reduce costs as well as improve the quality of the resulting data.

Cost reduction comes from simply collecting information for fewer units of analysis. In the typical public opinion poll, a sample of size 1,000 represents only 1 in 288,400 citizens of the United States. A representative sample of this size would allow a researcher to estimate the average age of an American with some precision, and the information could be collected more rapidly than if a CENSUS were taken of the entire population.

Some of the savings derived from using a sample can be applied to improving data quality. It could be applied to improving coverage in implementing the original sampling design by reducing NONRESPONSE. Additional resources could be used for INTERVIEWER TRAINING to improve the administration of the QUESTIONNAIRE. And resources could also be applied to lengthening the questionnaire so that more questions could be asked about specific concepts, improving the RELIABILITY and VALIDITY of their measurement.

Every sample is evaluated by two key properties: its design and its implementation. In order to invoke the SAMPLING DISTRIBUTION and employ

CONFIDENCE INTERVALS and confidence levels, a researcher must use a probability design—one in which every element of the population has a known, nonzero chance of being selected. Nonprobability designs such as QUOTA SAMPLES, CONVENIENCE SAMPLES, or PURPOSIVE SAMPLES do not permit this.

Every sample design must be well implemented by devoting resources to contacting every element in order to collect data for or from it. The coverage or response rate reflects the proportion of the original number of elements in the sample for which information has been obtained. Even if a probability design is used, a low coverage rate may invalidate the use of the sample estimates because of a concern that the loss of information was systematic and not RANDOM. In samples used by the government, for example, socioeconomic status or housing status often represent systematic sources of ERROR because low-income or homeless people may be difficult to find. In public opinion polls, individuals who are not interested in politics may be more likely to refuse to participate in TELEPHONE SURVEYS than those who are interested, so the results may reflect those with intensely held attitudes more than the total population.

—Michael W. Traugott

REFERENCES

- Kish, L. (1995). *Survey sampling*. New York: Wiley.
- Lavrakas, P. (1993). *Telephone survey methods: Sampling, selection, and supervision*. Newbury Park, CA: Sage.
- Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications*. New York: Wiley.

SAMPLING

The basic idea in sampling is **EXTRAPOLATION** from the part to the whole—from “the sample” to “the population.” (The population is sometimes rather mysteriously called “the universe.”) There is an immediate corollary: the sample must be chosen to fairly represent the population.

Methods for choosing samples are called “designs.” Good designs involve the use of probability methods, minimizing subjective judgment in the choice of units to survey. Samples drawn using probability methods are called “probability samples.”

BIAS is a serious problem in applied work; probability samples minimize bias. As it turns out, however, methods used to extrapolate from a probability sample to the population should take into account the method used to draw the sample; otherwise, bias may come in through the back door.

The ideas will be illustrated for sampling people or business records, but apply more broadly. There are sample surveys of buildings, farms, law cases, schools, trees, trade union locals, and many other populations.

SAMPLE DESIGN

Probability samples should be distinguished from “samples of convenience” (also called “grab samples”). A typical sample of convenience comprises the investigator’s students in an introductory course. A “mall sample” consists of the people willing to be interviewed on certain days at certain shopping centers. This too is a convenience sample. The reason for the nomenclature is apparent, and so is the downside: the sample may not represent any definable population larger than itself.

To draw a probability sample, we begin by identifying the population of interest. The next step is to create the “sampling frame,” a list of units to be sampled. One easy design is **SIMPLE RANDOM SAMPLING**. For instance, to draw a simple random sample of 100 units, choose one unit at random from the frame; put this unit into the sample; choose another unit at random from the remaining ones in the frame; and so forth. Keep going until 100 units have been chosen. At each step along the way, all units in the pool have the same chance of being chosen.

Simple random sampling is often practical for a population of business records, even when that population is large. When it comes to people, especially

when face-to-face interviews are to be conducted, simple random sampling is seldom feasible: where would we get the frame? More complex designs are therefore needed. If, for instance, we wanted to sample people in a city, we could list all the blocks in the city to create the frame, draw a simple random sample of blocks, and interview all people in housing units in the selected blocks. This is a “cluster sample,” the cluster being the block.

Notice that the population has to be defined rather carefully: it consists of the people living in housing units in the city, at the time the sample is taken. There are many variations. For example, one person in each household can be interviewed to get information on the whole household. Or, a person can be chosen at random within the household. The age of the respondent can be restricted; and so forth. If telephone interviews are to be conducted, **RANDOM DIGIT DIALING** often provides a reasonable approximation to simple random sampling—for the population with telephones.

CLASSIFICATION OF ERRORS

Since the sample is only part of the whole, extrapolation inevitably leads to **ERRORS**. These are of two kinds: sampling error (“random error”) and non-sampling error (“systematic **ERROR**”). The latter is often called “bias,” without connoting any prejudice.

SAMPLING ERROR results from the luck of the draw when choosing a sample: we get a few too many units of one kind, and not enough of another. The likely impact of sampling error is usually quantified using the “**SE**,” or **STANDARD ERROR**. With **PROBABILITY SAMPLES**, the **SE** can be estimated using (i) the sample design and (ii) the sample data.

As the “sample size” (the number of units in the sample) increases, the **SE** goes down, albeit rather slowly. If the population is relatively homogeneous, the **SE** will be small: the degree of heterogeneity can usually be estimated from sample data, using the standard deviation or some analogous statistic. Cluster samples—especially with large clusters—tend to have large **SEs**, although such designs are often cost-effective.

NON-SAMPLING ERROR is often the more serious problem in practical work, but it is harder to quantify and receives less attention than sampling error. Non-sampling error cannot be controlled by making the sample bigger. Indeed, bigger samples are harder to manage. Increasing the size of the

sample—which is beneficial from the perspective of sampling error—may be counter-productive from the perspective of non-sampling error.

Non-sampling error itself can be broken down into three main categories: (i) selection bias, (ii) non-response bias, and (iii) response bias. We discuss these in turn.

(i) **SELECTION BIAS** is a systematic tendency to exclude one kind of unit or another from the sample. With a convenience sample, selection bias is a major issue. With a well-designed probability sample, selection bias is minimal. That is the chief advantage of probability samples.

(ii) Generally, the people who hang up on you are different from the ones who are willing to be interviewed. This difference exemplifies non-response bias. Extrapolation from respondents to non-respondents is problematic, due to non-response bias. If the response rate is high (most interviews are completed), non-response bias is minimal. If the response rate is low, non-response bias is a problem that needs to be considered.

At the time of writing, US government surveys that accept any respondent in the household have response rates over 95%. The best face-to-face research surveys in the US, interviewing a randomly-selected adult in a household, get response rates over 80%. The best **TELEPHONE SURVEYS** get response rates approaching 60%. Many commercial surveys have much lower response rates, which is cause for concern.

(iii) Respondents can easily be led to shade the truth, by interviewer attitudes, the precise wording of questions, or even the juxtaposition of one question with another. These are typical sources of **RESPONSE BIAS**.

Sampling error is well-defined for probability samples. Can the concept be stretched to cover convenience samples? That is debatable (see below). Probability samples are expensive, but minimize selection bias, and provide a basis for estimating the likely impact of sampling error. Response bias and non-response bias affect probability samples as well as convenience samples.

TRADING NON-RESPONDENTS FOR RESPONDENTS

Many surveys have a planned sample size: if a non-respondent is encountered, a respondent is substituted.

That may be helpful in controlling sampling error, but makes no contribution whatsoever to reducing bias. If the survey is going to extrapolate from respondents to non-respondents, it is imperative to know how many non-respondents were encountered.

HOW BIG SHOULD THE SAMPLE BE?

There is no definitive statistical answer to this familiar question. Bigger samples have less sampling error. On the other hand, smaller samples may be easier to manage, and have less non-sampling error. Bigger samples are more expensive than smaller ones: generally, resource constraints will determine the sample size. If a pilot study is done, it may be possible to judge the implications of sample size for accuracy of final estimates.

The size of the population is seldom a determining factor, provided the focus is on relative errors. For example, the percentage breakdown of the popular vote in a US presidential election—with 200 million potential voters—can be estimated reasonably well by taking a sample of several thousand people. Of course, choosing a sample from 200 million people all across the US is a lot more work than sampling from a population of 200,000 concentrated in Boise, Idaho.

STRATIFICATION AND WEIGHTS

Often, the sampling frame will be partitioned into groupings called “strata,” with simple random samples drawn independently from each stratum. If the strata are relatively homogeneous, there is a gain in statistical efficiency. Other ideas of efficiency come into play as well. If we sample blocks in a city, some will be sparsely populated. To save interviewer time, it may be wise to sample such blocks at a lower rate than the densely-populated ones. If the objective is to study determinants of poverty, it may be advantageous to over-sample blocks in poorer neighborhoods.

If different strata are sampled at different rates, analytic procedures must take sampling rates into account. The “Horvitz-Thompson” estimator, for instance, weights each unit according to the inverse of its selection probability. This estimator is unbiased, although its variance may be high. Failure to use proper weights generally leads to bias, which may be large in some circumstances. (With convenience samples, there may not be a convincing way to control bias by using weights.)

An **ESTIMATOR** based on a complex design will often have a larger variance than the corresponding estimator

based on a simple random sample of the same size: clustering is one reason, variation in weights is another. The ratio of the two variances is called the DESIGN EFFECT.

RATIO AND DIFFERENCE ESTIMATORS

Suppose we have to audit a large population of claims to determine their total audited value, which will be compared to the “book value.” Auditing the whole population is too expensive, so we take a sample. A relatively large percentage of the value is likely to be in a small percentage of claims. Thus, we may over-sample the large claims and under-sample the small ones, adjusting later by use of weights. For the moment, however, let us consider a simple random sample.

Suppose we take the ratio of the total audited value in the sample claims to the total book value, then multiply by the total book value of the population. This is a “ratio estimator” for the total audited value of all claims in the population. Ratio estimators are biased, because their denominators are random: but the bias can be estimated from the data, and is usually offset by a reduction in sampling variability. Ratio estimators are widely used.

Less familiar is the “difference estimator.” In our claims example, we could take the difference between the audited value and book value for each sample claim. The sample average—dollars per claim—could then be multiplied by the total number of claims in the population. This estimator for the total difference between audited and book value is unbiased, and is often competitive with the ratio estimator.

Ratio estimators and the like depend on having additional information about the population being sampled. In our example, we need to know the number of claims in the population, and the book value for each; the audited value would be available only for the sample. For stratification, yet other information about the population would be needed. We might use the number of claims and their book value, for several different strata defined by size of claim. Stratification improves accuracy when there is relevant additional information about the population.

COMPUTING THE STANDARD ERROR

With simple random samples, the sample average is an unbiased estimate of the population

average—assuming that response bias and non-response bias are negligible. The SE for the sample average is generally well approximated by the SD of the sample, divided by the square root of the sample size.

With complex designs, there is no simple formula for variances; procedures like “the jackknife” may be used to get approximate variances. (The SE is the square root of the variance.) With non-linear statistics like ratio estimators, the “delta method” can be used.

THE SAMPLING DISTRIBUTION

We consider probability samples, setting aside response bias and non-response bias. An estimator takes different values for different samples (“sampling variability”); the probability of taking on any particular value can, at least in principle, be determined from the sample design. The probability distribution for the estimator is its “sampling distribution.” The expected value of the estimator is the center of its sampling distribution, and the SE is the spread. Technically, the “bias” in an estimator is the difference between its expected value and the true value of the estimand.

SOME EXAMPLES

In 1936, Franklin Delano Roosevelt ran for his second term, against Alf Landon. Most observers expected FDR to swamp Landon—but not the *Literary Digest*, which predicted that FDR would get only 43% of the popular vote. (In the election, FDR got 62%.) The *Digest* prediction was based on an enormous sample, with 2.4 million respondents. Sampling error was not the issue. The problem must then be non-sampling error, and to find its source, we need to consider how the sample was chosen.

The *Digest* mailed out 10 million questionnaires and got 2.4 million replies—leaving ample room for non-response bias. Moreover, the questionnaires were sent to people on mailing lists compiled from car ownership lists and telephone directories, among other sources. In 1936, cars and telephones were not as common as they are today, and the *Digest* mailing list was overloaded with people who could afford what were luxury goods in the depression era. That is selection bias.

We turn now to 1948, when the major polling organizations (including Gallup and Roper) tapped Dewey—rather than Truman—for the presidency. According to one celebrated headline,

DEWEY AS GOOD AS ELECTED, STATISTICS CONVINCE ROPER

The samples were large—tens of thousands of respondents. The issue was non-sampling error, the problem being with the method used to choose the samples. That was QUOTA SAMPLING. Interviewers were free to choose any subjects they liked, but certain numerical quotas were prescribed. For instance, one interviewer had to choose 7 men and 6 women; of the men, 4 had to be over 40 years of age; and so forth.

Quotas were set so that, in the aggregate, the sample closely resembled the population with respect to gender, age, and other control variables. But the issue was, who would vote for Dewey? Within each of the sample categories, some persons were more likely than others to vote Republican. No quota could be set on likely Republican voters, their number being unknown at the time of the survey. As it turns out, the interviewers preferred Republicans to Democrats—not only in 1948 but in all previous elections where the method had been used.

Interviewer preference for Republicans is another example of selection bias. In 1936, 1940, and 1948, Roosevelt won by substantial margins: selection bias in the polls did not affect predictions by enough to matter. But the 1948 election was a much closer contest, and selection bias tilted the balance in the polls. Quota sampling looks reasonable: it is still widely used. Since 1948, however, the advantages of probability sampling should be clear to all.

Our final example is a proposal to adjust the U. S. census. This is a complicated topic, but in brief, a special sample survey (“Post Enumeration Survey”) is done after the census, to estimate error rates in the census. If error rates can be estimated with sufficient accuracy, they can be corrected. The Post Enumeration Survey is a stratified block cluster sample, along the lines described above. Sample sizes are huge (700,000 people in 2000), and sampling error is under reasonable control. Non-sampling error, however, remains a problem—relative to the small errors in the census that need to be fixed. For discussion from various perspectives, see Imber (2001).

SUPER-POPULATION MODELS

Samples of convenience are often analyzed as if they were simple random samples from some large, poorly-defined parent population. This unsupported

assumption is sometimes called the “super-population model.” The frequency with which the assumption has been made in the past does not provide any justification for making it again, and neither does the grandiloquent name. Assumptions have consequences, and should only be made after careful consideration: the problem of induction is unlikely to be solved by fiat. For discussion, see Berk and Freedman (1995).

An SE for a convenience sample is best viewed as a de minimis error estimate: if this were—contrary to fact—a simple random sample, the uncertainty due to randomness would be something like the SE. However, the calculation should not be allowed to divert attention from non-sampling error, which remains the primary concern. (The SE measures sampling error, and generally ignores bias.)

SOME PRACTICAL ADVICE

Survey research is not easy; helpful advice will be found in the references below. Much attention needs to be paid in the design phase. The research hypotheses should be defined, together with the target population. If people are to be interviewed, the interviewers need to be trained and supervised. Quality control is essential. So is documentation. The survey instrument itself must be developed. Short, clear questions are needed; these should be worded so as to elicit truthful rather than pleasing answers. Doing one or more pilot studies is highly recommended. Non-response has to be minimized; if non-response is appreciable, a sample of non-respondents should be interviewed. To the maximum extent feasible, probability methods should be used to draw the sample.

—David A. Freedman

REFERENCES

- Berk, R. A., & Freedman, D. A. (1995). Statistical assumptions as empirical commitments. In T. G. Blomberg and S. Cohen (Eds.), *Law, punishment, and social control: Essays in honor of Sheldon Messinger* (pp. 245–258). New York: Aldine de Gruyter. 2nd ed. (2003) pp. 235–254.
- Cochran, W. G. (1977). *Sampling techniques*, 3rd ed. New York: John Wiley & Sons.
- Diamond, S. S. (2000). Reference guide on survey research. In *Reference manual on scientific evidence*, 2nd ed. (pp. 229–276). Washington, D. C.: Federal Judicial Center.
- Freedman, D. A., Pisani, R., & Purves, R. A. (1998). *Statistics*, 3rd ed. New York: W. W. Norton, Inc.

- Hansen, M. H., Hurwitz, W. N. & Madow, W. G. (1953). *Sample survey methods and theory*. New York: John Wiley & Sons.
- Imber, J. B. (Ed.). (2001). Symposium: Population politics. *Society*, 39, 3–53.
- Kish, L. (1987). *Statistical design for research*. New York: John Wiley & Sons.
- Sudman, S. (1976). *Applied sampling*. New York: Academic Press.
- Zeisel, H. & Kaye, D. H. (1997). *Prove it with figures*. New York: Springer.

SAMPLING BIAS. See SAMPLING ERROR

SAMPLING DISTRIBUTION

The sampling distribution of a particular statistic is the FREQUENCY DISTRIBUTION of possible values of the statistic over all possible samples under the sampling design. This distribution is important because it is equivalent to a PROBABILITY DISTRIBUTION, given that just one survey is usually carried out. Therefore, the single realized survey statistic is a RANDOM observation from the sampling distribution. Knowledge of the sampling distribution allows a researcher to calculate the probability that the statistic will fall within any specified range of values.

Two important characteristics of a sampling distribution are its location and its DISPERSION. If the mean of the DISTRIBUTION is located at the true population value of a PARAMETER, of which the sample statistic is an estimate, then the sample statistic is said to be an UNBIASED estimator of the population parameter. In other words, the estimate will be right “on average,” though not necessarily in any one realization. The greater the dispersion of the distribution, the more likely it is that a realized statistic—even if it is

an unbiased estimator—will be discrepant from the population parameter by any given magnitude. A common way to measure dispersion is by the variance of the distribution. Sample designers strive to minimize the variance of estimators because this minimizes the probability that the selected sample will produce an estimate considerably in error from the population parameter. In summary, sampling distributions of estimators should ideally be unbiased and have small variance.

Sampling distributions can be estimated only for designs that employ RANDOM SAMPLING. With non-probability designs, the concept of a probability distribution is not defined. With random sampling designs, estimators are known to be unbiased (under randomization inference), so only the variance of the sampling distribution remains of concern.

In practice, for a wide range of survey statistics, provided that the sample size is not too small, the sampling distribution turns out to approximate toward the NORMAL (Gaussian) DISTRIBUTION. In consequence, the known properties of the normal distribution can be used to provide approximate CONFIDENCE INTERVALS for estimates. For example, 95% of the possible estimates will lie within plus or minus 1.96 standard deviations of the mean of the distribution (Figure 1).

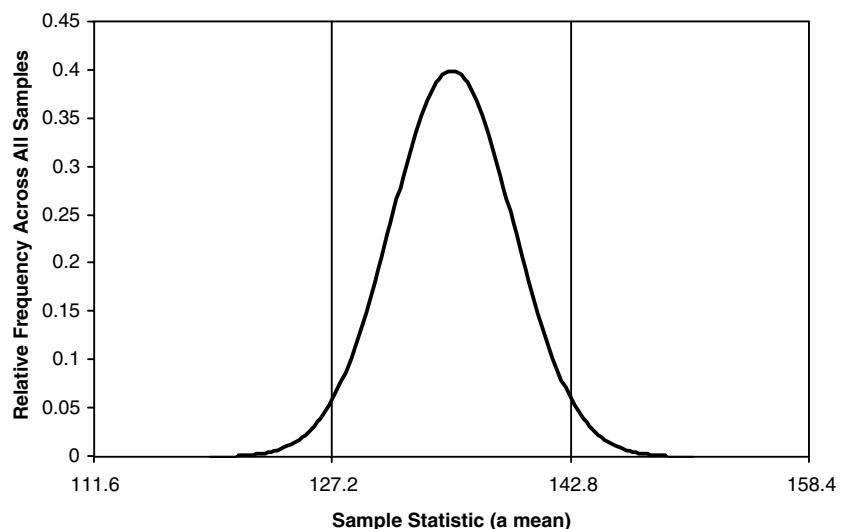


Figure 1 A Sampling Distribution

NOTES: The mean of the sampling distribution (expected value) is 135.0. The two vertical lines represent the mean plus or minus 1.96 standard deviations (the standard deviation is 3.98). Of all the possible estimates that could be obtained, 2.5% of them are less than 127.2, 2.5% are greater than 142.8, and the remaining 95% lie between these values.

Therefore, it is necessary only to estimate the STANDARD DEVIATION of the sampling distribution—more commonly referred to as the STANDARD ERROR of the estimate. However, the normal approximation is not universally good. In particular, certain types of estimates may tend to have a nonsymmetrical sampling distribution. In this situation, other methods must be used to estimate CONFIDENCE INTERVALS, such as replication methods.

It is important to realize that under a single sample design, different sample statistics will have different sampling distributions. Therefore, even if it is possible to estimate the mean and variance of sampling distributions prior to carrying out a survey, it will rarely be possible to select a design that is optimal for all statistics. Sample designs are typically a compromise, chosen to produce sampling distributions that are likely to be good enough for most of the estimates required.

—Peter Lynn

REFERENCES

- Mohr, L. B. (1990). *Understanding significance testing*. Newbury Park, CA: Sage.
- Moser, C. A., & Kalton, G. (1971). *Survey methods in social investigation* (2nd ed.). Aldershot, UK: Gower.
- Stuart, A. (1984). *The ideas of sampling*. London: Griffin.

SAMPLING ERROR

Any attempt to use a sample SURVEY as a means of studying a larger POPULATION of interest runs the risk that the selected sample will not have exactly the same characteristics as the population. In consequence, sample-based estimates will not necessarily take the same numeric values as the population PARAMETERS of which they are ESTIMATES. The term *sampling error* is used to refer to this general tendency of sample estimates to differ from the true population values. A main aim of sample design is to minimize the risk of sample characteristics differing greatly from population characteristics.

In the context of quantitative sample surveys using random sampling methods (see RANDOM SAMPLING), sampling error also has a more specific meaning. It refers to the expected deviation, under a particular specified sample design, of a particular sample

statistic from the equivalent population parameter. This expected deviation will have two components, a component due to random sampling error (variability, or VARIANCE) and a component due to systematic sampling error (BIAS). Therefore, for unbiased sample designs (or, strictly, unbiased sample-based estimators), sampling error is simply equal to sampling variance. Often, STANDARD ERRORS of survey estimates (the square root of the sampling variance) are estimated and presented as estimates of sampling error. Standard errors can be used in the construction of CONFIDENCE INTERVALS. More generally, MEAN SQUARED ERROR is used as a measure of sampling error. Mean squared error is defined as

$$\begin{aligned} \text{MSE}(y) &= E[y - E(y)]^2 + [E(y) - Y]^2 \\ &= \text{Var}(y) + \text{Bias}^2(y) \end{aligned} \quad (1)$$

where the sample statistic y is used to estimate the population parameter Y . The MSE is the sum of the variance and the squared bias of the estimate.

This quantity is, of course, specific to a particular estimate under a particular sample design. In practice, it can never be known for most survey estimates because the magnitude of bias is not known (if it was known, it could be removed from the survey estimate, so there would be no need to use a biased estimator). It is largely for this reason that attempts to estimate sampling error are typically restricted to estimation of the variance of an estimate. Also, there are many situations in which sampling bias can reasonably be assumed to be zero, or negligible. The variance of estimates can be readily estimated from survey data, provided that important design features such as clustering, stratification, and design weighting are indicated on the data and taken into account appropriately in the calculation. For most of the commonly used sample designs, the form of the variance is known; therefore, standard estimation methods can be used. For particularly complex designs, or designs that are not well specified, it may be necessary to use a REPLICATION method to estimate empirically the variance of sample statistics. Methods such as JACKKNIFE, BOOTSTRAP, and balanced repeated replications are increasingly available in commercial software packages.

Sampling error is only one component of total survey error. The total survey error of a statistic is the expected deviation of the statistic from the equivalent population parameter. This incorporates not only error due to sampling, but also error due to noncoverage;

NONRESPONSE; and MEASUREMENT (interviewer error, respondent error, instrument error, processing error). In many situations, although sampling bias may be negligible, other sources of bias may be important. For this reason, standard errors can be misleading as a guide to the accuracy of survey estimates.

—Peter Lynn

REFERENCES

- Scheaffer, R. L., Mendenhall, W., & Ott, L. (1990). *Elementary survey sampling*. Boston: PWS-Kent.
- Stuart, A. (1984). *The ideas of sampling*. London: Griffin.

very large (say, greater than .05)—leading to the often surprising result that a sample of 1,000, for example, is equally good for estimating the characteristics of a population of 100 million as it is for a population of 1 million.

—Peter Lynn

REFERENCES

- Barnett, V. (1974). *Elements of sampling theory*. Sevenoaks, UK: Hodder & Stoughton.
- Cochran, W. G. (1973). *Sampling techniques* (3rd ed.). New York: Wiley.
- Moser, C. A., & Kalton, G. (1971). *Survey methods in social investigation* (2nd ed.). Aldershot, UK: Gower.

SAMPLING FRACTION

The sampling fraction indicates the fraction, or proportion, of units on a SAMPLING FRAME that are selected to form the sample for a survey. If there are N units on the frame and n in the sample, then the sampling fraction is n/N .

Sampling fraction is sometimes confused with selection probability. A sampling fraction of n/N does not necessarily imply that each unit had a selection probability of n/N . It is possible that different units among the N had different selection probabilities. A sampling fraction simply indicates the overall fraction of population units that ended up in the sample, regardless of how they came to be selected.

Some surveys will use different sampling fractions for different parts of the POPULATION (strata). This is known as disproportionate stratified sampling. It is typically employed for the purpose of increasing the SAMPLE size of groups that are small in the population but of particular analytical interest. Using a larger sampling fraction in strata with larger population variance can also improve the precision of estimates (see also STRATIFIED SAMPLE).

In the case of systematic sampling, the sampling fraction reflects not only the outcome of the sampling process but also the process itself, and is equal in meaning to the selection probability. A sampling fraction of $1/A$ means that every A th unit on the list was (or will be) selected. A , the reciprocal of the sampling fraction, is referred to as the *sampling interval* (see also SYSTEMATIC SAMPLING).

The sampling fraction has a negligible effect on the precision of survey estimates unless it becomes

SAMPLING FRAME

A sampling frame is a list or other device used to define a researcher's POPULATION of interest. The sampling frame defines a set of elements from which a researcher can select a SAMPLE of the target population. Because a researcher rarely has direct access to the entire population of interest in social science research, a researcher must rely upon a sampling frame to represent all of the elements of the population of interest.

Generally, sampling frames can be divided into two types, list and nonlist. Examples of list frames include a list of registered voters in a town, residents listed in a local telephone directory, or a roster of students enrolled in a course. Examples of nonlist frames include a set of telephone numbers for RANDOM DIGIT DIALING and all housing units in a defined geographic area for area PROBABILITY SAMPLING. Random digit dial (RDD) frames are designed specifically for TELEPHONE SURVEY research, where interviewers call the households associated with selected telephone numbers to conduct an interview. An area probability sampling frame is primarily used only for in-person SURVEY research, where interviewers approach each sampled unit to complete an interview. Most list samples can be used with any data collection mode, including in-person, telephone, and mail surveys.

The two most important goals that a good sampling frame achieves are comprehensiveness and accuracy. Comprehensiveness refers to the degree to which a

sampling frame covers the entire target population. For example, a residential listing in a local telephone directory typically does not provide an adequate sampling frame to represent all residents in a given area. Telephone directories do not include residents with unlisted telephone numbers or those who have only recently begun telephone service. In contrast, an RDD sampling frame provides a current set of both listed and unlisted telephone numbers associated with a given set of telephone exchanges. Although the RDD frame is superior to the directory listing in this respect, neither frame includes residents in households without telephone service.

Accuracy refers to the degree to which a sampling frame includes correct information about the elements of the target population it covers. Telephone directories, town lists, and other kinds of lists often include information that is inaccurate because of data entry errors, outdated information, and other problems. The accuracy of the information provided by such lists can affect the quality of a sampling frame independent of its comprehensiveness.

Sampling frames that are not sufficiently comprehensive and accurate can be a major source of BIAS in any research project. Bias can occur when the sampling frame either excludes elements that are members of the target population or includes elements that are *not* members of the target population. To the extent that incorrectly excluded or included elements differ from the population of interest, these differences can produce bias. If a researcher can identify elements that have been incorrectly included in a sampling frame, the researcher can simply eliminate these elements from the sampling process. Yet researchers often have little information about elements excluded from the sampling frame that are members of the target population. Missing elements that cannot be identified and recovered have no chance of being selected into the sample. As a result, samples drawn from frames with missing elements often do not sufficiently represent the target population.

—Douglas B. Currivan

REFERENCES

- Fowler, F. J., Jr. (2002). *Survey research methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills, CA: Sage.

SAMPLING IN QUALITATIVE RESEARCH

SAMPLING is the deliberate selection of the most appropriate participants to be included in the study, according to the way that the theoretical needs of the study may be met by the characteristics of the participants. In qualitative inquiry, attention to sampling is crucial for the attainment of rigor and takes place throughout the research process.

TYPES OF SAMPLES

Samples are classified according to the means by which participants are selected from the population. In qualitative inquiry, this selection procedure is a *deliberate* rather than a *random* process.

Convenience Sample

Convenience sampling is selecting the sample by including participants who are readily available and who meet the study criteria. A CONVENIENCE SAMPLE may be used at the beginning of the sampling process, when the investigator does not know the pertinent characteristics for criteria for sample selection, or it is used when the number of participants available is small. For example, we may be exploring the experience of children who have had cardiac surgery, or of ventilator-dependent schoolchildren, and this patient population within a given geographic area is low. When *all* available participants have been included in the study (e.g., *all* children in a particular classroom), it is referred to as a *total* sample.

Minimally, to be included as a participant in a qualitative interview, participants must be willing and able to reflect on an experience, articulate their experience, and have time available without interruption to be interviewed. Problems may arise when using a total sample: When a participant is invited or volunteers to be interviewed, the researcher does not know if the participant will meet that minimal criterion for a “good” participant. If we include all participant interviews in the initial sample, we have not used any screening procedures. On the other hand, if the researcher starts an interview and finds that the participant is not a good informant and does not have the anticipated experiences nor knows the information we are seeking, then we practice *secondary selection* (Morse, 1989). In this

case, the researcher politely completes the interview. The tape is labeled and set aside. The researcher does not waste time and research funds TRANSCRIBING the tape. The interview is not incorporated into the data set, but neither is the tape erased. The audiotape is stored, just in case it makes sense and is needed at a later date.

Theoretical Sample

This is the selection of participants according to the theoretical needs of the study. If the researcher has identified a comparative research question and knows the dimensions of the comparison to be conducted, THEORETICAL SAMPLING may be used at the commencement of data collection. For example, the researcher may use a two-group design comparing participants—for instance, comparing single mothers who are continuously receiving unemployment with those who periodically work and sporadically claim benefits. Thus, theoretical sampling enables comparison of the two sets of data and the identification of differences between the two groups. As analysis proceeds, categories and themes form, some areas of analysis will be slower to saturate than others or may appear thin. A theoretical sample is the deliberate seeking of participants who have particular knowledge or characteristics that will contribute data to categories with inadequate data, referred to as *thin* areas of analysis. Alternatively, the investigator may develop a conjecture or HYPOTHESIS that could be tested by seeking out and interviewing people with characteristics that may be compared or contrasted with other participants. An example may be the findings from the previous study of single, unemployed mothers, which revealed that those mothers who returned to work were more resilient than those who remained unemployed. Because this sample had too few participants to verify this emerging finding, the researchers must then deliberately seek out participants who fit that criterion, in order to saturate those categories. This type of theoretical sampling, in which participants are deliberately sought according to individual characteristics, is sometimes also referred to as PURPOSIVE SAMPLING.

Negative Cases

Sampling for negative cases is the deliberate seeking of participants who are exceptions to the developing theoretical scheme. If we carry our unemployed single mother example further, a negative case, such as a

mother who returned to work periodically, but who does *not* demonstrate the identified traits of resilience, may emerge. In order for the developing theory to be complete and valid, negative cases must be purposefully selected until the negative category is saturated, and logical explanation for these exceptions included in the theory.

THE PROGRESSION OF QUALITATIVE SAMPLING

As noted, in qualitative inquiry, sampling continues throughout the study, forming a responsive link between data collection and analysis. Initially, the investigator may practice *open sampling*—deliberately seeking as much variation as possible within the limits of the topic. This variation enables the scope of the problem or the phenomena to be delineated. Later, as sampling proceeds, as understanding is gained, and as categories begin to emerge, the sampling strategies may change to theoretical sampling in order to verify emerging theoretical formulations. Thus, sampling continues until the theory is thick, dense, and saturated.

Thus far, we have referred to sampling as the process of selecting participants for the study. In qualitative inquiry, sampling also occurs during the process of analysis, in the selection of and attending to particular data. In qualitative analysis, all data are not treated equally—the researcher attends to some data, data that may be richer or provide more explanation that contributes to the developing insights. Other data may be used to support or refute these developing notions, and yet still other data may not contain any relevant information, and therefore may not be attended to in a particular phase of analysis and consequently ignored. At the writing-up stage, the researcher purposely selects the best examples of quotations to illustrate points made in the developing argument.

WHY RANDOM SAMPLES ARE INAPPROPRIATE

Randomization means that anyone in the identified group is a potential participant in the study. Yet qualitative researchers know that all participants would not make equally good participants. We know that some participants may be more reluctant to participate than others (they may not have time, may not want to think about the topic, may feel ill, and so forth), and some may be more or less articulate and

reflective than others. Of greatest concern, not all have equal life experiences that are of interest to the interviewer. Because qualitative data are expensive to obtain—interviews are very time consuming and must be transcribed; text is clumsy and time consuming to analyze, compared with test scores and quantitative analysis—qualitative researchers cannot afford to collect data that are weak or off the topic. In addition, the process of randomization does not give researchers the opportunity to control the useful amount of data they collect. Because knowledge of a topic is not randomly distributed in the population, collecting data by chance would result in excessive data being collected on topics that are generally known, and inadequate data on less common topics. Thus, the data contained will be excessive in some areas (and therefore waste resources) and inadequate in others. Furthermore, the common findings, that is, the amount of data in topics that occur frequently, may overwhelm and mute certain less common and equally important features, thus actually biasing and invalidating the analysis.

EVALUATION OF THE SAMPLE

Those new to qualitative research have a number of concerns: How many participants? How many observations? How many data will be needed? How many data depends on a number of factors, including the quality of the data, the nature of the phenomenon, and the scope of the study. Consider the following principles:

1. *The broader the scope of the study, the greater the amount of data that is needed.* A common trap for a new investigator is to not delineate the problem before beginning the study, so that he or she tries to study everything—which, of course, means that many more participants and many more data will be needed to reach saturation. On the other hand, prematurely delineating the topic may result in excluding important aspects of the problem or simplifying the problems, and be equally invalid.

2. *The better the quality of the interviews, the fewer interviews that will be needed.* Ideally, interviews should be conducted with an appropriate sample. If the sample is carefully selected, and if the information obtained is minimally irrelevant, is insightful, and has excellent detail, then the investigator will be able to stop sampling sooner than if he or she obtains interviews that are vague, irrelevant, ambiguous, or contradictory.

3. *There is an inverse relationship between the amount of data obtained from each participant and the number of participants.* Methods that use in-depth interviewing methods such as interactive interviews, which involve interviewing participants three times for more than 1 hour each, will reach saturation more quickly, with fewer participants, than if they used SEMISTRUCTURED INTERVIEWS and presented participants with only a set number of question stems.

HOW DO YOU TELL WHEN YOU HAVE SAMPLED ENOUGH?

There is an adage that the qualitative researcher may stop sampling when he or she is bored (Morse & Richards, 2002). That or, saturation is reached when no new data appear, and when the researcher has “heard it all.” Saturation occurs more quickly when the sample is adequate and appropriate, as well as when the study has a narrow focus.

Adequacy

Adequacy refers to enough data. Data must replicate within the data set. That is, it is not satisfactory to hear something from someone once—these ideas should be verified by other participants, or verified by other means (documents, observations, and so forth). These verifications may be in the same form or in different forms—for instance, different stories that illustrate similar points that the researcher is trying to make. Adequate data are essential for concept and theory constructions—in fact, it is harder to develop categories with thin, inadequate data than with adequate data, for it is difficult to see patterns without REPLICATION, and themes if there are gaps.

Appropriateness

Appropriateness refers to the deliberate selection of the best participants to be involved in the study. Appropriate sampling in qualitative inquiry is ongoing throughout the study, using the methods of purposeful and theoretical sampling described above.

—Janice M. Morse

REFERENCES

- Morse, J. M. (1989). Strategies for sampling. In J. Morse (Ed.), *Qualitative nursing research: A contemporary dialogue* (pp. 117–131). Rockville, MD: Aspen.

Morse, J. M., & Richards, L. (2002). *Read me first for a qualitative guide to qualitative methods*. Thousand Oaks, CA: Sage.

SAMPLING VARIABILITY. See **SAMPLING ERROR**

SAMPLING WITH (OR WITHOUT) REPLACEMENT

SIMPLE RANDOM SAMPLES are, by convention, samples drawn without replacement. **SAMPLING** without replacement means that when a unit is selected from the population to be included in the sample, it is not placed back into the pool from which the sample is being selected. In sampling without replacement, each unit in the population can be selected for the sample once and only once. For example, if we are drawing a sample of 10 balls from a group of 999 balls, we could place the balls in a box, each numbered with a unique number between 1 and 999. Then, the box is shaken and the first ball selected in a fair process. If we are drawing a sample of 10 balls, the number on the first one—in this example, 071—is noted and the ball is set aside. Nine other balls are selected using the same process, and each time, the one that is selected is set aside. Because the balls are set aside after being selected for the sample, this is sampling without replacement. By convention, sampling without replacement is assumed in most samples taken for the purposes of social science research.

Although this procedure may seem to be common sense, from the standpoint that it seems to serve no purpose for a member of the study population to appear more than once in a single sample, an alternative procedure, sampling with replacement, has some very nice properties and is actually very important for statistical theory. Sampling with replacement, using the same shaken container of balls as the earlier example, would involve selecting each ball—this time, say, the first one is 802—recording its number, and placing the ball back into the box. This time, 802 is eligible for selection again. A benefit from the standpoint of statistical theory is that each draw for selecting a sample member is completely **INDEPENDENT** of the selection of any other

unit. Each population member can be selected on any draw, thus, the **PROBABILITY** of selection is the same for each draw.

For sampling without replacement, each draw after the first draw is dependent on the prior draws. Although the dependence of future selections on previous draws is relatively minor when selecting **PROBABILITY SAMPLES** without replacement, some issues should be considered. For example, an issue of independence arises in the selection of a **CONTROL GROUP** in randomized **EXPERIMENTS**. If the **TREATMENT** group is selected first, then, from a technical standpoint, the selection of the sample for the control group is dependent on the selection of the treatment group. If the randomization procedures are not completely without **BIAS**, as happened with the lottery for the draft into the US military in 1969 (McKean, 1987, p. 75), this can create a bias. In most cases, the occurrence of such bias can be investigated by examining the differences between the treatment and control groups on characteristics that are related, according to theory, to the outcome variables. The differences can show whether the bias is likely to affect the experiment's results, but can never be disproved with certainty.

Another concern with practical implications is the calculation of the **STANDARD ERROR** for samples without replacement. Technically speaking, when a sample without replacement is drawn, a finite population control (fpc) is required to compute the standard error for the sample. The formula for a simple random sample is

$$s_{(e)} = (1 - f)^{1/2} s / n^{1/2},$$

where

$s_{(e)}$	is the standard error of estimator e
$1 - f$	is the finite population correction factor
f	is the sampling fraction, or n/N
s	is the estimate of the standard deviation
n	is the size of the sample

The influence of the fpc depends upon the size of the population and the size of the sample relative to the population. Generally, the fpc is of no practical significance when the sample is less than 5% of the population (Henry, 1990; Kish, 1965). If, for example, a sample of 1,200 is to be drawn from a population of 10,000 students in a school district, the fpc would reduce the sampling error by approximately .94, a meaningful amount. If the same-size sample was being drawn from a population of 1.4 million students in an

entire state, the fpc would reduce the sampling error by .9996, an insignificant amount that effectively makes the fpc = 1. In most cases for social science research, the fpc can be omitted, but when drawing a sample from a small population, it can significantly reduce the sampling error for samples drawn without replacement.

—Gary T. Henry

REFERENCES

- Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.
- Kish, L. (1965). *Survey sampling*. New York: Wiley.
- McKean, K. (1987, January). The orderly pursuit of pure disorder. *Discover*, pp. 72–81.

SAS

SAS (Statistical Analysis System) is a general-purpose software package that is used by a wide variety of disciplines, including the social sciences. For further information, see the Web site of the software: <http://www.sas.com/software/index.html>.

—Tim Futing Liao

SATURATED MODEL

A saturated model has as many parameters as data values. The predicted or fitted values from a saturated model are equal to the data values; that is, a saturated model perfectly reproduces the data. A single observation or data point can be thought of as one piece of information; therefore, in a sample of n data points, there are n pieces of information. Each statistic or parameter computed from n data points uses up one piece of information. When a model has as many parameters as there is information in the data, the model is saturated, and the model is said to have zero DEGREES OF FREEDOM.

As a small example, suppose that we have a sample of three observations— $x_1 = 3$, $x_2 = 5$, and $x_3 = 10$ —from some population, and we want to fit the model $x_i = \mu + \beta_i$ to the data where μ and β_i are parameters of the model. The β_i s may have a different value for each observation (i.e., $i = 1, 2$, and 3). A reasonable estimate of μ is the sample mean; that is,

$\bar{x} = (3 + 5 + 10)/3 = 6$, which uses up one piece of information and leaves two pieces of information to estimate the β_i s. With the remaining two pieces of information, we can estimate two β_i s by taking the difference between a data value and the sample mean. For example, $\hat{\beta}_1 = (x_1 - \bar{x}) = (3 - 6) = -3$, and $\hat{\beta}_2 = (x_2 - \bar{x}) = (5 - 6) = -1$. We now have three estimated parameters, $\bar{x}$, $\hat{\beta}_1$, and $\hat{\beta}_2$, which uses up all the information in the data. The value of the last parameter, β_3 , is computed using $\hat{\beta}_1$ and $\hat{\beta}_2$, along with the fact that the sum of differences between the data values and their mean equals zero; that is,

$$\begin{aligned}\sum_{i=1}^3 (x_i - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) \\ &= \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = 0.\end{aligned}$$

Solving this for $\hat{\beta}_3$ gives us $\hat{\beta}_3 = -(\hat{\beta}_1 + \hat{\beta}_2) = -(-3 - 1) = 4$. Using the estimated parameters in the model yields predicted values that are exactly equal to the data values; that is,

$$\begin{aligned}\hat{x}_1 &= 6 - 3 = 3 \\ \hat{x}_2 &= 6 - 1 = 5 \\ \hat{x}_3 &= 6 + 4 = 10.\end{aligned}$$

Models are fitted to data to provide a summary or description of data, to produce an equation for predicting future values, or to test hypotheses about the population from which the data were sampled. Generally, simpler models are preferred because they smooth out the unsystematic variation present in data, which results from having only a sample from a population. Although saturated models are extremely accurate, they do not distinguish between unsystematic and systematic variation in data.

Saturated models play an important role in global GOODNESS-OF-FIT statistical tests for nonsaturated models. All nonsaturated models are special cases of the saturated model. Consider two common statistics for testing model goodness-of-fit to data: PEARSON'S CHI-SQUARE STATISTIC,

$$\chi^2 = \sum_{i=1}^N (x_i - \hat{x}_i)^2 / \hat{x}_i,$$

and the LIKELIHOOD RATIO STATISTIC,

$$L^2 = \sum_{i=1}^N x_i \log(x_i / \hat{x}_i),$$

where x_i equals an observed measurement on individual i , $\hat{x}_i$ equals the predicted or modeled measurement from an unsaturated model, and N equals the total number of observations. In both cases, the nonsaturated model's predicted values are compared to data values, which are the predicted values of the saturated model.

—Carolyn J. Anderson

REFERENCES

- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.
- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.
- Powers, D. A., & Xie, Y. (2000). *Statistical methods for categorical data*. San Diego, CA: Academic Press.

SCALE

A scale is a composite measure of some underlying CONCEPT. Many phenomena in social science are theoretical CONSTRUCTS that cannot be directly measured by a single VARIABLE. To understand the causes, effects, and implications of these phenomena, the researcher must develop a VALID and RELIABLE empirical indicator of these constructs. This empirical indicator is called a *scale*. In this sense, the scale is composed of a set of measurable items that empirically captures the essential meaning of the theoretical construct.

Good scales are data reduction devices that simplify the information and, in one composite, measure the direction and intensity of a construct. They are usually constructed at the ORDINAL level of measurement, although some can be at the INTERVAL level. Because they are constructed of more than one indicator of a particular phenomenon, they increase both the reliability and the validity that would be attainable if the individual indicators were used in analysis.

Several possible models can be used to combine items into a scale. These include THURSTONE scaling, LIKERT scaling, and GUTTMAN scaling. The method chosen depends on the purpose that the scale is intended to serve.

SCALING has three related but distinct purposes. In the first case, scaling may be intended to test a specific HYPOTHESIS (e.g., that a single dimension, party identification, structures voters' choice of presidential candidate). In this case, the scaling model is used as

a criterion to evaluate the relative fit of a given set of observed data to a specific model.

Another purpose for scaling is to describe a data structure. If a political scientist attempted to discover the underlying dimensions (e.g., social, economic, cultural, etc.) of the United States' electoral system, this would be an exploratory analysis rather than a hypothesis-testing approach.

The third purpose for scaling is to construct a measure on which individuals can be placed and then relate their scores on that scale to other measures of interest.

Scaling models may be used to scale people, stimuli, or both people and stimuli. The Likert scale is a scaling model that scales only subjects. Broadly, any SCALE obtained by summing the response scores of its constituent items is called a Likert or "summative" scale, or a linear composite. An examination in a mathematics class is an example of a linear composite. The scale (test) score is found by adding the number of correct answers to the individual items (questions). The composite score on the scale (test) is a better indicator of the student's knowledge of the material than is any single item (question).

Thurstone's interest lay in measuring and comparing stimuli when there is no evident logical structure. The Thurstone scaling model is an attempt to identify an empirical structure among the stimuli. To do this, Thurstone scaling uses human judgments. Individuals are given statements, maybe as many as 100. These "judges" order the statements along an underlying dimension. Those statements that produced the most agreement among the judges are selected as the items to be included in the scale. These remaining items (say, 20) should cover the entire latent continuum. These items are then presented to subjects, who are asked to identify those they accept or with which they agree. The average scale value of those items chosen represents the person's attitude toward the object in question.

The final type of scale to be discussed here, Guttman scaling (or scalogram analysis or cumulative scaling), is a procedure designed to order both items and subjects with respect to some underlying cumulative dimension. It was developed to be an empirical test of the unidimensionality of a set of items. The underlying logic of this method is cumulative; that is, by knowing a subject's score on the scale, that subject's response to each item is known. If the subject's score is 3, the researcher knows not only that the subject answered three items "correctly," but also which three, namely,

the first, second, and third items. Much of Guttman scaling analysis concerns what to do with deviations from the perfect scale.

This has been a cursory discussion of scales—what they are and the purposes for which they are used. Scales are used extensively in social research, and their construction and interpretation bears some study.

—Edward G. Carmines and James Woods

REFERENCES

- Carmines, E. G., & McIver, J. P. (1981). *Unidimensional scaling* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-024). Beverly Hills, CA: Sage.
- Guttman, L. L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139–150.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 44–53.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273–286.
- Thurstone, L. L. (1929). Fechner's law and the method of equal-appearing intervals. *Journal of Experimental Psychology*, 12, 214–224.
- Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitudes*. Chicago: University of Chicago Press.

SCALING

Scaling methods comprise procedures that assign numbers to properties or characteristics of objects. As such, “scaling” seems to be very similar to “measurement” (which is also commonly defined as “assigning numbers to objects”). The difference between the two is subtle, but important: MEASUREMENT can usually be reduced to some kind of physical counting operation. For example, the width of an object is assessed by determining the number of uniformly sized rods (or uniformly spaced calibration marks on a single rod) that can be placed alongside the object, spanning the entire distance from one side to the other. Similarly, the mass of an object is determined by the number of uniform weights that are required to balance it in a scale.

GENERAL OBJECTIVES

Scaling, in contrast, is used to quantify phenomena that cannot be counted directly. For example,

scaling techniques have been employed to measure the attitudes of survey respondents, the severity of crimes, the prestige of occupations, and the ideology of legislators. In each of these cases, we believe that the phenomenon in question (i.e., attitudes, severity, prestige, ideology) actually exists. But none of these things can be “counted” in a manner similar to the way one would assess the calibrations on a ruler or a scale. It is for such phenomena that scaling methods exist.

There is no single technique called “scaling.” Instead, there is a very wide variety of scaling models. The common characteristic is that they all seek to recover systematic structure in real-world observations by representing elements of the input data matrix (i.e., observations, stimuli, variables, etc.) as objects (usually, points or vectors) within a geometric space. Thus, a scale is an abstract model of data that are, themselves, culled from empirical observations. The way that one proceeds from observations, to data, to the geometric model depends upon the specific scaling strategy.

In any case, performing a scaling analysis is similar to putting together a jigsaw puzzle. Like the pieces of a puzzle, individual data points provide only fragmentary, limited information. When puzzle pieces are put together, the result is a picture that an observer can comprehend and understand. Similarly, when observations are combined into some geometric structure, the resultant “picture” shows the relative positions of the objects in the data matrix and thereby helps the analyst understand the contents and implications of the data themselves.

DATA AND SCALING METHODS

How does one choose a specific scaling model for a particular data analysis situation? The answer depends upon the researcher's interpretation of information contained within the input data matrix and the type of model that he or she wants to construct to represent that information. The two most common types of data matrices are *multivariate* and *similarities*. Multivariate data refer to a matrix in which the rows and columns represent different objects (usually, observations and variables), and the cell entries provide scores for one set of objects with respect to the other (observations' variable values). For example, survey respondents' ratings of the incumbent president on each of 10 characteristics would be an example of multivariate data.

With similarities data, the rows and columns represent the same objects (hence, the matrix is actually square), and the cell entries indicate the degree to which the row and column entries are similar to each other: for example, an experimental subject's opinion about the degree to which each of 10 foods tastes similar to each of the others. But similarities data can represent many other things besides actual similarity among objects. For scaling purposes, such diverse phenomena as distances, covariances, and the degree to which different objects share common characteristics can all be interpreted as similarities data. Probably the most familiar type of similarities data is a CORRELATION matrix (which measures the pairwise "similarity" of variables).

If the researcher has multivariate data and wants to obtain a unidimensional scale of the rows *or* the columns, then LIKERT SCALING (also called SUMMATED RATING SCALING) would be appropriate. If the objective is a unidimensional representation of the rows *and* columns, then GUTTMAN SCALING (based upon the cumulative scaling model) would be a useful technique. If the analyst is interested in locating the rows and columns within multiple dimensions, then PRINCIPAL COMPONENTS ANALYSIS could be used. Turning to similarities data, if the input is a correlation MATRIX, then FACTOR ANALYSIS is often the appropriate model: Correlations are modeled as functions of the angles between vectors (which represent the variables). MULTIDIMENSIONAL SCALING (MDS) is used for most other kinds of similarities data: In that case, similarities are modeled as distances between points in a space (with greater similarity corresponding to smaller distance).

Although less common, there are many other types of data besides multivariate and similarities—and scaling methods appropriate for them. For example, a data matrix in which the rows and columns represent the same objects and each cell entry shows the degree to which that row object exceeds the column object (or vice versa) on some specified characteristic is sometimes called *stimulus comparison* data. This kind of information can be scaled using any of several THURSTONE SCALING techniques. A matrix in which the rows and columns represent different objects, with cell entries that indicate the proximity of each row element to each column element, is sometimes called *preferential choice* data. It can be used as input to an unfolding analysis.

DATA REDUCTION

Researchers tend to employ scaling procedures for four complementary reasons. First, there is data reduction. Analysts are often confronted with massive amounts of information, often too much to be readily comprehensible. Therefore, scaling methods can be used to "distill" the information and re-express its most essential components in a more useable form. In so doing, the resultant scale often produces a summary measure that is superior to any of the constituent variables that comprise it. This occurs because the strengths of one item in the scale can compensate for the weaknesses of another. Thus, the effects of measurement error are minimized.

For example, a public opinion survey might ask its respondents several questions about their feelings toward the incumbent president. The responses to these questions could then be combined into a single scale of attitudes toward the president. The motivations for doing so are that (a) a single variable is easier to analyze than several variables, and (b) the scale would presumably provide a more accurate representation of each respondents' feelings about the president than would any of the individual questions.

DIMENSIONALITY

Second, scaling methods can be used to assess the dimensionality of data. From a substantive perspective, "dimensionality" refers to the number of separate and interesting sources of variation that exist among the objects under investigation. From an analytic perspective, dimensionality refers to the number of coordinates that are required to provide an accurate map of the geometric elements (again, usually points or vectors) used to represent those objects. Again, all scaling procedures seek to represent objects within a space. The dimensions of the space correspond to the sources of variability underlying the objects.

Sometimes, the sources of variability among objects are known beforehand. Therefore, the researcher can specify the dimensionality of the space before the task of locating the object points is begun. This is the case for physical measurement as well as for any other context where objects are compared to each other on the basis of predetermined criteria. For example, an arithmetic test can be viewed as a scaling analysis that seeks to locate a set of objects (students) along a single dimension (corresponding to mathematical skill).

In other situations, scaling procedures are intended to *identify* the number of dimensions that are required. This is particularly important in the social and behavioral sciences, because researchers often do not have a priori knowledge about the underlying sources of variation in their observations. For example, can the various political systems of the world be readily differentiated according to a single continuum that separates democratic from authoritarian systems? Or, would doing so oversimplify a more complex set of distinctions between the sociopolitical structures that actually exist in the world?

In a scaling analysis, dimensionality is assessed by examining the goodness of fit for a particular scaling solution. The actual fit measures differ across the various techniques. For example, with a summated rating scale, the RELIABILITY COEFFICIENT is used to assess the fit. In factor analysis, the sum of the EIGENVALUES or communalities can be used as an overall fit measure. In nonmetric multidimensional scaling, fit is assessed by the stress coefficient. Regardless of the precise details, goodness of fit usually means the degree to which the original data can be predicted from the components of the abstract model that is constructed by the scaling technique. Dimensionality is usually defined by the scaling solution with the smallest number of dimensions that still exhibits an adequate fit to the data.

MEASUREMENT

The third reason for employing scaling methods is measurement. From a substantive perspective, this means obtaining numerical estimates of the relative positions of the objects along the relevant dimensions of variability. In terms of the scaling model, measurement involves finding the coordinates for the scaled objects (again, points or vectors) within the geometric space recovered by the scaling procedure. Once this is accomplished, the numeric values (or “scale scores”) can be used as empirical variables for subsequent analyses. When employed in this manner, scaling methods provide operational representations of phenomena that are inherently unobservable.

For example, summed item scores for responses to a set of survey questions about the president create a Likert scale that serves to measure the respondents’ attitudes toward the president. Alternatively, an MDS analysis might require subjects to assess the similarities among a set of occupations. In the resultant MDS solution, the coordinates of the points could be

interpreted as the occupations’ positions with respect to the various cognitive criteria that the subjects used to evaluate the occupational similarities. In still another application, a researcher might factor analyze data on state governments’ spending across a specified set of public policies. Factor scores could be calculated for the states and interpreted as the priorities that the respective states assign to the policy areas corresponding to the different factors. In every case, the critical point is that the scale(s) produced by the analysis is assumed to provide better measurement of the phenomenon under investigation than any of the original variables.

When scaling solutions are used for measurement purposes, there are several issues that should be confronted. For one thing, the researcher must confirm the VALIDITY of the scale scores as empirical representations of the property that he or she intends to measure. This is a thorny problem, precisely because scales are often employed to measure unobservable phenomena; hence, determining the consistency between the two can be quite difficult. The usual approach is to examine the scale relative to separate, external criteria—either alternative measures of the same property or other variables with theoretically predictable relationships to the property that the scale is measuring. If the scale scores show an appropriate pattern of empirical correlations with other variables, then it is strong evidence in favor of the researcher’s interpretation. If the predicted relationships do not occur, then the scale may be confounding some alternative source of variation in the data, apart from the property that the analyst is trying to measure.

Another problem with some scaling strategies is that the scaling solution is consistent with a sizable (and often infinite) number of different dimensional arrays of the objects under investigation. The final results of a factor analysis or multidimensional scaling analysis locate vectors or points (respectively) within a space of specified dimensionality. However, the orientations of the coordinate axes that define the space are usually arbitrary. If the vectors or points are held at rigid positions with respect to each other, then the axes can be rotated around the configuration to any other location. Of course, changing the orientation of an axis also changes the vector/point coordinates on that axis. This is problematic if the latter are interpreted in substantive terms because the values of the scale scores would change accordingly. With most scaling procedures, there is no way around this problem. Therefore,

the researcher must justify the final scaling solution in terms of factors that are separate from the scaling technique itself (e.g., parsimonious interpretation, fidelity to substantive theory, consistency with external criteria).

GRAPHICAL PRESENTATION

The fourth reason for performing a scaling analysis concerns the presentation of the final results: Because scaling methods seek to produce geometric models of data within a low-dimensional space, it is often possible to construct a graphical depiction of the resultant scaling solution. For example, factor analysis produces a space (hopefully, with few dimensions) containing vectors that represent the common variation within the variables being factor analyzed. Similarly, multidimensional scaling produces a point configuration such that the distances between points correspond to measured dissimilarities between the objects contained in the input data matrix. But even inherently unidimensional methods (e.g., Likert or Guttman scales) can be shown graphically. One of the reasons for using such methods is that the scale provides more detailed resolution of the property being investigated than do the individual items. Hence, a histogram of the scale scores provides far more information about the distribution of objects with respect to the relevant property than would a histogram of scores on any individual item.

The graphical displays obtained from most scaling analyses are very useful for teasing out the substantive implications of the analysis. Interpretation is a creative act on the part of the researcher, accomplished by searching for meaningful systematic structures within the geometric model created by the scaling technique. It is often convenient to do this through visual inspection of the results, because patterned regularities are usually discerned more easily in graphical, rather than numeric, information.

CONCLUSIONS

In summary, scaling methodology provides strategies for modeling systematic structure that exists within a data matrix. Such methods are particularly appropriate when the structure is believed to be due to the existence of latent, or unobservable, sources of variation in the data. The latter feature is precisely what makes scaling so important to the social and behavioral sciences: The central concepts in substantive theories

are inherently unobservable; scaling methods provide tools for operationalization. However, scales only produce models, which are never “right” or “wrong.” They merely represent the analyst’s theory about the nature of the processes that generated the empirical observations. Scaling analyses can generate theory-relevant insights and useful empirical variables. But like any other theory, they remain tentative statements about the nature of reality, subject to challenge if a different scaling model generates an equally parsimonious, interpretable, and analytically useful geometric representation of the same data.

—William G. Jacoby

REFERENCES

- Coombs, C. H. (1964). *A theory of data*. New York: Wiley.
- Jacoby, W. G. (1991). *Data theory and dimensional analysis*. Newbury Park, CA: Sage.
- Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.
- Young, F. W., & Hamer, R. M. (1987). *Multidimensional scaling: History, theory, and applications*. Hillsdale, NJ: Lawrence Erlbaum.

SCATTERPLOT

A scatterplot is a two-dimensional coordinate graph showing the RELATIONSHIP between two QUANTITATIVE VARIABLES, with each observation in a data set plotted as a point in the graph. Although they are employed less frequently in publications than some other statistical graphs, scatterplots are arguably the most important graphs for data analysis. Scatterplots are used not only to graph variables directly, but also to examine derived quantities, such as RESIDUALS from a REGRESSION MODEL.

Precursors to scatterplots date at least to the 17th century, such as Pierre de Fermat’s and René Descartes’s development of the coordinate plane, and Galileo Galilei’s studies of motion. The first statistical use of scatterplots is probably from Sir Francis Galton in the 19th century, and is closely tied to his pioneering work on CORRELATION, regression, and the bivariate-NORMAL DISTRIBUTION.

Several versions of an illustrative scatterplot in Figure 1 show the relationship between fertility and contraceptive use in 50 developing countries. The total fertility rate is the number of expected live births to

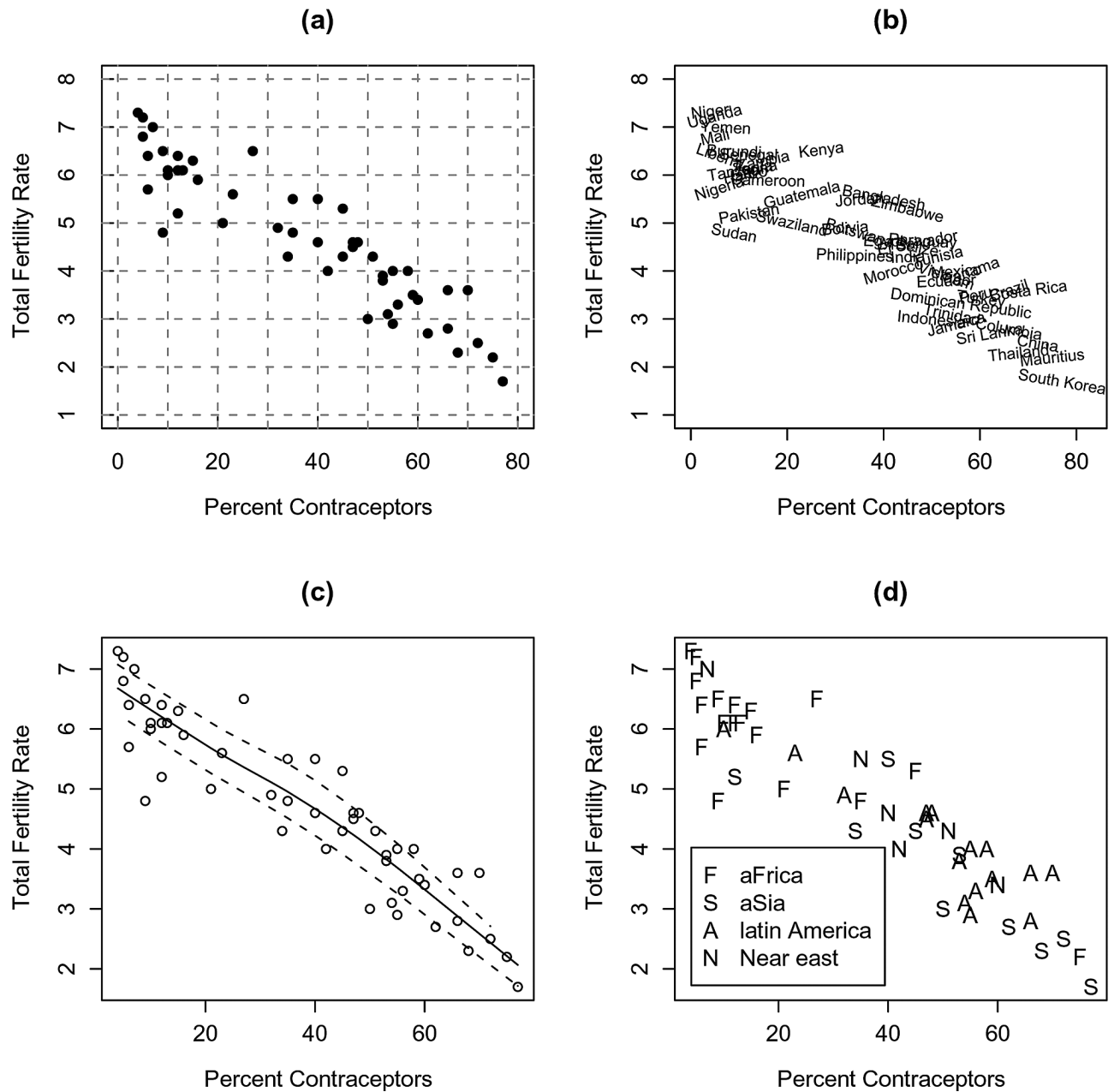


Figure 1 Four Versions of a Scatterplot of the Total Fertility Rate (Number of Expected Births per Woman) Versus the Percentage of Contraceptors Among Married Women of Childbearing Age, for 50 Developing Nations

SOURCE: Robey, Rutstein, and Morris (1992).

a woman surviving her child-bearing years at current age-specific levels of fertility. As is conventional, the explanatory variable (contraceptive use) appears on the horizontal axis, and the response variable (fertility) on the vertical axis.

- A traditional scatterplot is shown in Panel A.
- In Panel B, the names of the countries are graphed in place of points. The names are drawn at random angles to minimize overplotting, but many are still hard to read; if the graph were enlarged without

proportionally enlarging the plotted text, this problem would be decreased.

- Panel C displays a more modern version of the scatterplot: Grid lines are suppressed; tick marks do not project into the graph; the data are plotted as open, rather than filled, circles to make it easier to discern partially overplotted points; and the axes are drawn just to enclose the data. The solid curve in the graph represents a ROBUST NONPARAMETRIC REGRESSION smooth; the two broken curves are smoothes based on the positive and negative residuals from the nonparametric regression, and are robust estimates of the STANDARD DEVIATION of the residuals in each direction. The relationship between fertility and contraceptive use is nearly LINEAR; the residuals have near-constant spread and appear symmetrically distributed around the regression curve. More information on smoothing scatterplots may be found in Fox (2000).

- Panel D illustrates how the values of a third, categorical variable may be encoded on a scatterplot. Here, region is encoded using discriminable letters. Similar encodings employ different colors or plotting symbols.

Beyond the illustrations in Figure 1, there are a number of important variations on and extensions of the simple BIVARIATE scatterplot: Three-dimensional dynamic scatterplots may be rotated on a computer screen to examine the relationship among three variables; scatterplot matrices show all pairwise marginal relationships among a set of variables; and interactive statistical graphics permit the identification of points on a scatterplot, and the linking of scatterplots to one another and to other graphical displays. These and other strategies are discussed in Cleveland (1993) and Jacoby (1997).

—John Fox

REFERENCES

- Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart.
- Fox, J. (2000). *Nonparametric simple regression: Smoothing scatterplots*. Thousand Oaks, CA: Sage.
- Jacoby, W. G. (1997). *Statistical graphics for univariate and bivariate data*. Thousand Oaks, CA: Sage.
- Robey, B., Rutstein, S. O., & Morris, L. (1992). The reproductive revolution: New survey findings (Technical Report

M-11). *Population Reports*. Baltimore, MD: Johns Hopkins University Center for Communication Programs.

SCHEFFÉ'S TEST

Henry Scheffé (1953) developed this test to determine which means or combination of means differed significantly from which other means or combination of means in an ANALYSIS OF VARIANCE (in which the scores of the groups are unrelated) when the overall *F* RATIO for those means in that analysis was significant and when such differences had not been predicted prior to the analysis. This type of test is more generally known as a *post hoc* or a *posteriori* test, or a multiple comparison test. It takes into account the fact that a significant difference is more likely to occur by chance the more comparisons that are made. It is suitable for groups of unequal size. Compared to other post hoc tests, it tends to be more conservative in the sense that a significant difference is less likely to be obtained by chance (Toothaker, 1991).

In an analysis of variance, a significant *F* ratio for a factor of more than two groups or for an interaction between two or more factors shows that the variances between the groups differ significantly from the variance expected by chance but does not show which means or combination of means differ significantly. If there were strong grounds for predicting particular differences, the statistical significance of these differences could be determined with an a priori test, such as a one-tailed unrelated *T*-TEST. If there were no strong grounds for predicting particular differences, the number of possible comparisons that can be made is generally greater. For example, the number of means that can be compared two at a time can be calculated with the following formula: $[J \times (J - 1)]/2$, where *J* represents the number of means. So, with 4 means, the number of possible pairwise comparisons is $6 - [4 \times (4 - 1)]/2 = 6$. If the probability of finding a significant difference is set at the usual 0.05 level, the probability of finding one or more of these pairwise comparisons significant at this level can be worked out with the following formula: $1 - (1 - 0.05)^J$. This probability level is known as the familywise error rate, and for 4 comparisons, it is about $0.17 - [1 - (1 - 0.05)^4] = 0.17$. However, there are a number of other comparisons that can be made apart from the pairwise ones that

increase the overall number of potential comparisons. For example, the means of the first two groups can be combined and compared with that of the third or fourth group or the third and fourth groups combined.

The Scheffé test controls for the familywise error rate for more than pairwise comparisons by multiplying the critical value of the F ratio for a factor or interaction by 1 less than the number of groups to be compared. For 4 groups, with 5 cases in each, the .05 critical value of the F ratio is about 3.24, as the degrees of freedom are 3 for the numerator and 16 for the denominator. Consequently, the critical value of the F ratio for the Scheffé test for any comparison of the means of these 4 groups is about 9.72 ($3 \times 3.24 = 9.72$). If the F ratio for the analysis of variance is statistically significant, then at least one of the comparisons of the means must also be statistically significant.

—Duncan Cramer

REFERENCES

- Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, 40, 87–104.
- Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park, CA: Sage.

SCREE PLOT

Raymond Cattell (1966) developed the scree test as an alternative to methods such as the Kaiser-Guttman test for determining the number of factors in a FACTOR ANALYSIS that explain a nontrivial amount of the variance that is shared between or common to the variables. It is the first point at which the amount of variance explained by a factor appears to be very similar to that explained by subsequent factors, and it generally tends to be small, suggesting that these factors can be ignored. Scree is a geological term for the rubble and boulders that lie at the base of a steep slope and obscure the real base of the slope itself. In this sense, the test is used to determine which factors can be considered as scree or debris and which do not contribute substantially to any underlying structure of the variables.

In the scree test, the amount of the variance that is accounted for by the factors is plotted on a graph such as that shown in Figure 1. The vertical axis of the graph represents the amount of variance accounted for by a factor. The horizontal axis shows the number of the factor arranged in order of decreasing size of amount of variance, which is the order in which they emerge. The

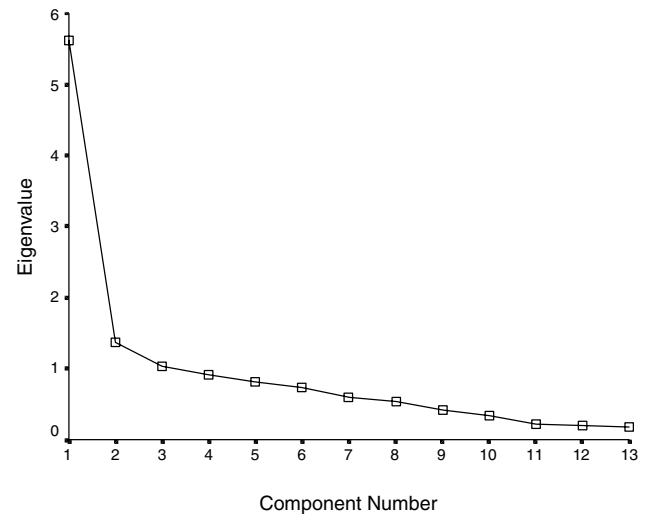


Figure 1 A Scree Plot

amount of variance explained by a factor is often called the *eigenvalue* or *latent root*. The eigenvalue is the sum of the squared loading or correlation of each variable with that factor. Figure 1 shows the EIGENVALUES of a principal components analysis of 13 variables, which results in 13 factors or components. The scree appears to begin at component number 3. In other words, this test suggests that 2 of the 13 factors should be retained. The point at which the scree begins may be determined by drawing a straight line through or very close to the points represented by the scree, as shown in Figure 1. The factors above the first line represent the number of factors to be retained.

A factor or principal components analysis will always give as many factors or components as there are variables. However, the aim of factor analysis is to determine whether the variables can be grouped together into a smaller number of aggregated or super variables called *factors*. The first factor will always explain the greatest amount of variance between the factors. Subsequent factors will explain decreasing amounts of variance. A criterion needs to be established to determine how many of the subsequent factors can be ignored because they explain too little variance. The Kaiser-Guttman test decrees that factors with eigenvalues of 1 or less should be ignored because these factors explain, at most, no more than the variance of a single variable. In Cattell's experience, this criterion resulted in too few factors with few variables and too many factors with many variables, although he did not specify to what numbers these adjectives referred. Zwick and Velicer (1986) suggest that with 36 and

72 variables, this criterion overestimated the number of factors to a greater extent than did the scree test.

—Duncan Cramer

REFERENCES

- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245–276.
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432–442.

SECONDARY ANALYSIS OF QUALITATIVE DATA

Also called *qualitative secondary analysis*, secondary analysis of qualitative data is the reexamination of one or more existing qualitatively derived data sets in order to pursue research questions that are distinct from those of the original inquiries. Because qualitative inquiries often involve intensive data collection using methods such as SEMISTRUCTURED INTERVIEWS, PARTICIPANT OBSERVATION, and FIELDWORK approaches, they typically create data sets that contain a wealth of information beyond that which can be included in a primary research report. Furthermore, with the fullness of time, new questions often arise for which existing data sets may be an efficient and appropriate source of grounded knowledge.

Secondary analysis also provides a mechanism by which researchers working in similar fields can combine data sets in order to answer questions of a comparative nature or to examine themes whose meaning is obscured in the context of smaller samples. When data sets from different populations, geographical locations, periods of time, or situational contexts are subjected to a secondary inquiry, it becomes possible to more accurately represent a phenomenon of interest in a manner that accounts for the implications of the original inquiry approaches. For example, Angst and Deatrick (1996) were able to draw our attention to discrepancies between chronically ill children's involvement in everyday as opposed to major treatment decisions by conducting a comparative secondary analysis on data sets from their respective prior studies of children living with cystic fibrosis and those about to undergo surgery for scoliosis.

Although QUALITATIVE RESEARCHERS have often created multiple reports on the basis of a single data set,

the articulation of secondary analysis as a distinctive qualitative research approach is a relatively recent phenomenon (Heaton, 1998). Methodological precision and transparency are critically important to the epistemological coherence and credibility of a secondary analysis in order to account for the implications of the theoretical and methodological frame of reference by which the qualitative researcher entered primary study (Thorne, 1994). Regardless of the number of data sets and the level of involvement the secondary researcher may have had in creating them, a secondary analysis inherently involves distinct approaches to conceptualizing sampling, data collection procedures, and interaction between data collection and analysis (Hinds, Vogel, & Clarke-Steffen, 1997).

Qualitative secondary analysis can be an efficient and effective way to pose questions that extend beyond the scope of any individual research team, examine phenomena that would not have been considered relevant at the original time of inquiry, or explore themes whose theoretical potential emerges only when researchers within a field of study examine each others' findings.

—Sally E. Thorne

REFERENCES

- Angst, D. B., & Deatrick, J. A. (1996). Involvement in health care decisions: Parents and children with chronic illness. *Journal of Family Nursing, 2*(2), 174–194.
- Heaton, J. (1998). Secondary analysis of qualitative data [Online]. *Social Research Update, 22*. Retrieved April 28, 2002, from <http://www.soc.surrey.ac.uk/sru/SRU22.html>.
- Hinds, P. S., Vogel, R. J., & Clarke-Steffen, L. (1997). The possibilities and pitfalls of doing a secondary analysis of a qualitative data set. *Qualitative Health Research, 7*, 408–424.
- Thorne, S. (1994). Secondary analysis in qualitative research: Issues and implications. In J. Morse (Ed.), *Critical issues in qualitative research methods* (pp. 263–279). Thousand Oaks, CA: Sage.

SECONDARY ANALYSIS OF QUANTITATIVE DATA

Secondary analysis refers to the analysis of DATA originally collected by another researcher, often for a different purpose. This article focuses specifically on secondary analysis of QUANTITATIVE data, primarily from SURVEYS and CENSUSES. Various definitions, as well as the range of data sources that may be candidates for secondary analysis, are discussed by Hyman

(1972), Hakim (1982), and Dale, Arber, and Procter (1988).

Secondary analysis of high-quality data sets can provide an extremely cost-effective means by which researchers, often with only limited funding, can reap the benefits of a large investment for very little additional cost. The time-consuming and expensive processes of planning and conducting a survey, and cleaning and documenting the data are avoided entirely. However, because the data have been collected by another researcher, the secondary analyst will have had no opportunity to influence the questions asked or the CODING FRAMES used, and this important factor must be borne in mind at all stages of analysis.

HISTORICAL DEVELOPMENT

Secondary analysis of quantitative data became possible with the development of the computer in the 1950s and 1960s. Computerization meant that data could be stored in a form that was compact and readily transportable, and, of great importance, the data could be anonymized so that they could be used by others without revealing identification. The development of computer packages such as SPSS for statistical analysis opened the way for easier and more sophisticated analysis of survey data by many more people than had ever before been possible.

As a response to these developments, DATA ARCHIVES were established for the deposition, cataloging, storage, and dissemination of machine-readable data. During the 1960s, archives were established in the United States (Roper Public Opinion Research Center and the Inter-university Consortium for Political and Social Research, University of Michigan); Europe; and the United Kingdom (UK Data Archives, University of Essex). Since this time, the number of users, as well as the number of studies deposited, has expanded steadily. Also, the type of data has broadened in scope to include administrative records and aggregate data such as area-level statistics from censuses.

APPLICATION

In both Britain and the United States, the availability of large-scale government surveys has been an additional factor of considerable importance in the development of secondary analysis. Government surveys provide data with particular features that are very difficult to obtain through other routes. They are usually nationally representative; with large sample

sizes; conducted by personal interview; and, of great importance, carried out by permanent staff with considerable expertise in SAMPLING, questionnaire design, FIELDWORK, and all other aspects of survey work. The net result is that these surveys provide data of a range and quality that are well beyond the funding that can be obtained by most academic researchers. Secondary analysis has specific value in the following areas.

National-Level Generalizations Across a Wide Range of Topics

Secondary analysis can provide the basis for making generalizations to the POPULATION as a whole. Large government surveys are particularly important because they often cover a wide range of topics in considerable depth and have large enough sample numbers to provide estimates with a high level of accuracy.

Historical Comparisons

Secondary analysis may be the only means by which historical comparisons can be made, particularly where information cannot be collected retrospectively—for example, income and expenditure, or attitudes. Data archives allow the researcher to go back in time and find sources of information on what people thought, how they voted, and how much they earned, perhaps 20 years ago. Many government surveys began in the early 1970s or before, and support analyses of change over time in, for example, women's economic role within the family over a 20-year period.

Longitudinal Analysis

The value of LONGITUDINAL data in testing HYPOTHESES and establishing CAUSAL processes is increasingly recognized. However, the difficulties of establishing and maintaining a longitudinal study for more than a few years are such that secondary analysis is, for most researchers, the only option. In Britain, many longitudinal studies—for example, the British Household Panel Survey—are specifically designed as national resources. In the United States, longitudinal studies such as the Panel Study of Income Dynamics and the National Longitudinal Study of Youth provide similarly rich resources for academic research.

International Comparisons

International comparisons are of great research importance but pose enormous problems of cost and organization and provide another example where secondary analysis plays a vital role. Sometimes, it is

possible to locate data sources from different countries with sufficient similarity in the topics and questions asked to support comparative research—for example, the social mobility studies between England, France, and Sweden (Erikson, Goldthorpe, & Portocarero, 1979). In other cases, surveys are *designed* to allow international comparisons—for example, the European Social Survey, which is designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs, and behavior patterns of its populations.

Representative Data on Small Groups

Secondary analysis can also provide a means of obtaining data on small groups for whom there is no obvious SAMPLING FRAME. Many population surveys will have a sufficiently large sample to support the analysis of relatively small subgroups. For example, a study of the elderly would require considerable screening in order to locate an appropriate sample but can be readily conducted by secondary analysis. Samples of microdata from censuses are typically large enough to support analyses of minority ethnic groups (Dale, Fieldhouse, & Holdsworth, 2000).

Contextual Data for a Qualitative Study

It is often valuable to incorporate a nationally representative dimension into a QUALITATIVE study in order to establish the size and distribution of the phenomenon under study. For example, a qualitative study concerned with expenditure decisions by single parents might be considerably enhanced by including background information that located single parents and their level of income and expenditure within the population as a whole. Often, this kind of contextual information can be obtained cheaply and fairly easily through secondary analysis, in circumstances where the cost of a nationally representative survey would be out of the question. The use of secondary analysis to identify the social and demographic characteristics of the group of interest may also be helpful to qualitative researchers when selecting their sample.

Conversely, secondary analysis can highlight interesting relationships that need to be pursued using qualitative methods.

Confidentiality of Data

Although CONFIDENTIALITY of data is a fundamental consideration in all social research, it is particularly

important when data collected for a specific purpose are made available for analysis by others. Standard precautions are always taken to ensure that individuals cannot be identified, including anonymizing the data by ensuring that name and address are not attached, restricting the geographical area to a level that will not allow identification, and suppressing any peculiar or unusual characteristics that would allow a respondent to be uniquely identified.

—Angela Dale

REFERENCES

- Dale, A., Arber, S., & Procter, P. (1988). *Doing secondary analysis*. London: Unwin Hyman.
- Dale, A., Fieldhouse, E., & Holdsworth, C. (2000). *Analyzing census microdata*. London: Arnold.
- Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational mobility in three Western European societies. *British Journal of Sociology*, 30, 415–441.
- Hakim, C. (1982). *Secondary analysis of social research*. London: George Allen and Unwin.
- Hyman, H. H. (1972). *Secondary analysis of sample surveys*. Glencoe, IL: Free Press.

SECONDARY ANALYSIS OF SURVEY DATA

Secondary analysis of survey data is the reanalysis of existing survey responses with research questions that differ from those of the original research. With an increasing number of data sets available to researchers, the role of surveys in social research has expanded, and, with improved data quality and access, secondary analysis has assumed an increasingly central position. Secondary survey data are available on a wide range of topics covering health, attitudes, population characteristics, criminal behavior, aging, and more, and many existing data sets are useful for both cross-sectional and temporal analyses. A major advantage of secondary survey analysis is resource economy. Both time and money are saved, and the need for extensive personnel is eliminated. Because the data already exist, the analysis is typically noninvasive, although discussions regarding human subjects implications are ongoing. The availability of multiple data sets simplifies comparative analyses—most often of nations or over time.

The limitations of this method are usually those of VALIDITY, relating to the appropriateness of existing

data for particular research questions. Because the secondary researcher did not design the original survey, it is unlikely that the data set will contain all variables of interest. Even those that do appear may have been measured in ways that diminish their utility. With existing data sets, researchers must be certain that samples represent the population of interest or have enough cases of the needed subpopulation that can be easily extracted. The UNIT OF ANALYSIS must also be compatible with the study. Although many existing data sets were developed with a specific focus, there has been some effort to promote omnibus surveys, that is, those covering a wide range of topics. The most widely known of these is the General Social Survey, which is conducted by the National Opinion Research Center (NORC) at the University of Chicago and contains an array of social indicator data.

Secondary survey data are now widely available on disk, CD-ROM, and the Internet. Many data repositories exist—some with a huge number of holdings covering a variety of topics, and others that are smaller and more specialized. Although academic archives, government agencies, and public opinion research centers (such as Roper and Gallup) are obvious sources of secondary survey data, any organization that stores such data (e.g., private research firms, foundations, and charitable organizations) can be considered a resource for the secondary analyst. A leading data source for academic social science researchers is the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan, which can be accessed on the Internet at <http://www.icpsr.umich.edu>. The ICPSR is an excellent starting point for locating desired data. In addition to their substantial holdings, the Web site provides links to other archives.

To make the best use of existing survey data, researchers should merge general substantive interests with familiarity of existing data and be able to evaluate primary data source(s) with respect to survey design and administration, issues of sampling, and data set limitations. Furthermore, researchers must be creative in their approach to combining data both within and across surveys, and by incorporating data from outside the data set. Although inherent challenges persist in the use of existing survey data, the opportunities for social research are enormous.

—Laura Nathan

SECONDARY DATA

Analysts may gather DATA themselves by carrying out a SURVEY, doing an EXPERIMENT, CONTENT ANALYZING newspapers, or CODING observed behaviors, to name a few methods, all of which generate primary data. In contrast, secondary data are data gathered by someone other than the analyst. A great deal of social science work is done on secondary data. For example, a social researcher may go to the library and gather together data first assembled by the U.S. Census Bureau. Or, a psychologist may collect data on a number of published experiments testing a particular hypothesis, in order to carry out a META-ANALYSIS of those studies. Another example is the use of data archived in data banks, such as the Inter-university Consortium for Political and Social Research. Increasingly, secondary data are available on the Web. Secondary data have the advantage that they are much less costly than primary data because you do not have to gather them yourself. They have the disadvantage that they may not contain the particular OBSERVATIONS or VARIABLES that you would have included if you had directed the study. Increasingly, general social surveys (such as the General Social Survey) are conducted without a particular primary research goal for each collection, but instead to facilitate researchers using them as what might be called “secondary” primary data.

—Michael S. Lewis-Beck

SECOND-ORDER. See ORDER

SEEMINGLY UNRELATED REGRESSION

Seemingly Unrelated Regression or Seemingly Unrelated Regression Equations (SURE) is a method used when two or more separate REGRESSION equations that have different sets of INDEPENDENT VARIABLES are related to each other through correlated DISTURBANCE TERMS. Estimation with SURE produces more efficient coefficients than estimation with separate regressions, especially when the disturbances are highly correlated and the sets of independent variables are not highly

correlated. The SURE estimation method can be applied to pooled TIME-SERIES CROSS-SECTION data that are either time-series dominant or cross-section dominant. It can also be applied to CROSS-SECTIONAL DATA and to TIME-SERIES DATA. If the error terms of the equations are not correlated, then ORDINARY LEAST SQUARES (OLS) REGRESSION can be used on each equation separately.

A model of federal agency budgets illustrates the SURE method. If the DEPENDENT VARIABLE is spending levels for federal agencies, and if separate models are estimated for different types of agencies, then the disturbance terms of the separate models are likely to be correlated. That is, the spending levels for the different types of agencies are all likely to be influenced by common exogenous shocks. For example, the U.S. war on terrorism produces large demands on the budget that influence the spending levels of various nondefense agencies.

SURE is also best applied when each equation in the set of equations has a different set of independent variables. In the budget example, this allows the researcher to recognize that not all agencies are alike and to group the agencies by type in different models. Grouping agencies by type and putting them into separate equations allows for agencies to be influenced differently by the same independent variables and also to be influenced by different independent variables. For example, unemployment may influence regulatory agency budgets like the Nuclear Regulatory Agency differently from how it influences the budget for the Employment and Training Administration under the Department of Labor. And unemployment may have no influence at all on regulatory agencies like the Federal Communications Commission, which means it should not be in the model for regulatory agencies. In contrast, OLS or GENERALIZED LEAST SQUARES (GLS) estimation of pooled agency data assumes that all of the agencies are influenced the same way by the same independent variables. That is, not only are the data pooled together, but the estimates are also pooled together.

—Charles Tien

REFERENCES

- Kmenta, J. (1986). *Elements of econometrics* (2nd ed.). New York: Macmillan.
- Pindyck, R., & Rubinfeld, D. (1991). *Econometric models and economic forecasts* (3rd ed.). New York: McGraw-Hill.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *American Statistical Association Journal*, 57, 348–368.

SELECTION BIAS

Selection bias is an important concern in any social science research design because its presence generally leads to inaccurate estimates. Selection bias occurs when the presence of observations in the SAMPLE depends on the value of the variable of interest. When this happens, the sample is no longer randomly drawn from the population being studied, and any inferences about that population that are based on the selected sample will be biased. Although researchers should take care to design studies in ways that mitigate nonrandom selection, in many cases, the problem is unavoidable, particularly if the data-generating process is out of their control. For this reason, then, many methods have been developed that attempt to correct for the problems associated with selection bias. In general, these approaches involve modeling the selection process and then controlling for it when evaluating the outcome variable.

For an example of the consequences of selection on INFERENCE, consider the effect of SAT scores on college grades. Under the assumption that higher SAT scores are related to success in college, one would expect to find a strong and positive relationship between the two among students admitted to a top university. The students with lower SAT scores who are admitted, however, must have had some other strong qualities to gain admittance. Therefore, these students will achieve unusually high grades compared to other students with similar SAT scores who were not admitted. This will bias the estimated effect of SAT scores on grades toward zero. Note that the selection bias does not occur because the sample is overrepresentative of students with high SAT scores, but because the students with low scores are not representative of all students with low SAT scores in that they score high on some unmeasured component that is associated with success.

A frequent misperception of selection bias is that it can be dismissed as a concern if one is merely trying to make inferences about the sample being analyzed rather than the population of interest, which would be akin to claiming that we have correctly estimated the

relationship between SAT scores and grades among admittees only. This is false: The grades of the students in our example are representative of neither all college applicants nor students with similar SAT scores. Put another way, our sample is not a random sample of students conditional on their SAT scores. The estimated relationship is therefore inaccurate. In this example, the regression coefficient will be biased toward zero because of the selection process. In general, however, it may be biased in either direction, and in more complicated (multivariable) regressions, the direction of the bias may not be known.

Although selection bias can take many different forms, it is generally separated into two different types: CENSORING and TRUNCATION. Censoring occurs when all of the characteristics of each observation are observed, but the dependent variable of interest is missing. Truncation occurs when the characteristics of individuals that would be censored are also not observed; all observations are either complete or completely missing. In general, censoring is easier to deal with because data exist that can be used to directly estimate the selection process. Models with truncation rely more on assumptions about how the selection process is related to the process affecting the outcome variable.

The most well-known model for self-selectivity is James Heckman's ESTIMATOR for a continuous dependent variable of interest that suffers from censoring (Heckman, 1979). This estimator corrects for selection bias in the linear regression context and has been widely used in the social sciences. Heckman's correction attempts to eliminate selection bias by modeling the selection process using a PROBIT equation where the dependent variable is an indicator for whether or not the observation is censored. It then estimates the conditional expected value of the error term in the equation of interest, which is not zero if it is correlated with the error in the selection equation. This variable, known as the inverse Mills' ratio, is then included as an additional regressor in the equation of interest, which is estimated using OLS. The standard errors need to be adjusted if the model is estimated this way because they understate the uncertainty in the model. Alternatively, both equations can be estimated simultaneously using maximum likelihood techniques, and no adjustment is needed. When its assumptions are met, Heckman's estimator provides unbiased, consistent estimates and

also provides an estimate of the correlation between the two error terms.

The presence of selection bias can be tested for in the Heckman model by conducting a significance test on the coefficient for the selection term or, more generally, by conducting a LIKELIHOOD RATIO TEST for the fit of the selection model relative to the uncorrected model. Care must be taken in interpreting the effects of independent variables on the dependent variable of interest because some will influence it indirectly through the selection equation as well as directly in the equation of interest. Another important consideration is the presence of factors that influence that selection equation that do not directly influence the quantity of interest. Although the model is identified (by functional form alone) without such factors, the inclusion of factors that belong in the selection equation but not the equation of interest will generally improve the reliability of the model because they increase the variation of the inverse Mills' ratio.

More advanced techniques have been developed for more complicated settings, including those where the data are truncated, which makes it impossible to directly estimate a selection equation, and those where the dependent variable of interest is binary rather than continuous (Dubin & Rivers, 1989). For an overview of these methods, see Maddala (1983); for more advanced approaches, including semiparametric methods that relax distributional assumptions, see Winship and Mare (1992). In general, the more advanced methods have been developed in response to findings that standard selection bias models are often sensitive to the assumptions of the correction, particularly that the two error terms are distributed bivariate normal.

—Frederick J. Boehmke

REFERENCES

- Dubin, J. A., & Rivers, D. (1989). Selection bias in linear regression, logit and probit models. *Sociological Methods & Research*, 18, 360–390.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- Maddala, G. G. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.
- Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology*, 18, 327–350.

SELF-ADMINISTERED QUESTIONNAIRE

Self-administered QUESTIONNAIRES are one of the most frequently used methods for collecting DATA for research studies. Furthermore, self-administered questionnaires appear in many areas of our lives. Think, for example, of the testing strategies used in most classrooms from kindergarten through graduate school. Classroom tests are a type of self-administered questionnaire. Similarly, we fill out forms or questionnaires to obtain everything from a driver's license to a death certificate.

The very proliferation and familiarity of self-administered questionnaires in a wide variety of daily life settings often results in people assuming that they can develop a questionnaire literally overnight and use it to collect data immediately. But, like any research endeavor, the development of self-administered questionnaires takes time and thought if the researcher wants to maximize the collection of complete, reliable, and valid data.

TYPES OF SELF-ADMINISTERED QUESTIONNAIRES

Questionnaires can be completed in either supervised or unsupervised settings. In supervised settings, someone is available to answer questions that respondents might have, monitor whether respondents talk with each other while filling out the questionnaire, and encourage respondents to complete the questionnaire and return it to the researcher or other personnel. Passing out questionnaires in a classroom is an example of a supervised group administration. Questionnaires may also be given to individuals to fill out in less structured settings such as registration lines, where a data collector is available but less obvious.

Questionnaires that are mailed to respondents are the most common example of unsupervised administration, but increasingly, questionnaires are being administered through e-mail, the Internet, and other ONLINE METHODS. Self-administered questionnaires used in all settings must stand alone, or be completely self-explanatory, but this is particularly important when unsupervised administrative procedures are used.

ADVANTAGES AND DISADVANTAGES

The single greatest advantage of self-administered questionnaires is their lower cost compared to other data collection methods. MAIL QUESTIONNAIRES have three sample-related advantages—wider geographic coverage, larger samples, and wider coverage within a sample population—and all self-administered questionnaires are easier to implement because fewer personnel are needed for data collection, processing, and analysis. Unlike most other methods of data collection, it can be assumed that when a questionnaire is sent through the mail or distributed online, all members of the sample receive it nearly simultaneously. Many researchers believe that more complete and truthful information on SENSITIVE TOPICS is obtained with self-administered questionnaires.

Mail and self-administered questionnaires have a number of disadvantages that limit or prohibit their use in many research settings. Address lists must be available if mail or online administration is to be used. One of the biggest and most studied disadvantages of mail questionnaires is the low response rate. When no effort is made to follow up the original mailing with additional mailings or phone contacts, the expected response rate in a general community sample is around 20%. To date, online survey response rates appear to be even lower. Self-administered questionnaires cannot be used with the estimated 20% of the population that is functionally illiterate, or with populations that include substantial proportions of respondents who do not read the language in which the research is being conducted. Elderly people or other groups with visual problems may also find it difficult or impossible to fill out questionnaires.

Self-administered questionnaires can be used only when the objective of the study is clear and not complex. The need for a clear and noncomplex data collection objective has ramifications for how the questionnaire is constructed and precludes the use of many strategies typically used in designing questionnaires used in interviews. First, questionnaires must be shorter and, hence, the number of questions asked and the amount of data collected are reduced. In general, the upper limit for a self-administered questionnaire is thought to be 12 pages, but length varies with respondent motivation. When a topic is particularly important to respondents, they will tolerate a longer questionnaire. Second, the questions must be CLOSED-ENDED. Although highly motivated respondents may

be willing to answer a few OPEN-ENDED QUESTIONS, the researcher who develops a questionnaire dominated by open-ended questions will find that few questionnaires will be completed and returned, and that those that are returned will have substantial amounts of MISSING or irrelevant data. Third, the questionnaire must stand alone. In other words, all the information that the potential respondent needs to answer the questions must be provided on the questionnaire itself because there is no interviewer available to clarify instructions or provide additional information. Fourth, in simplifying the task for respondents, the surveyor cannot use branches or skips. Every question should contain a response category that each respondent can comfortably use to describe his or her attitudes, behavior, knowledge, or characteristics. In some instances, this means that a “not applicable” alternative must be included among the responses provided.

The single biggest administrative disadvantage with unsupervised self-administered questionnaires is the fact that once the questionnaire leaves the researcher’s office, he or she has no control over who, in fact, completes it and whether that person consults with others when completing it. Furthermore, the researcher has no control over the order in which questions are answered. As a result, self-administered questionnaires cannot be used when one set of questions is likely to contaminate, BIAS, or influence answers to another set of questions.

DESIGNING SELF-ADMINISTERED QUESTIONNAIRES

Three kinds of information must be evaluated in deciding whether or not data can be collected by mail or other self-administered questionnaire. These are the literacy level and motivation of the targeted population, and whether characteristics of the RESEARCH QUESTION make it amenable to data collection using a self-administered questionnaire. Literacy is usually fairly easily assessed, but motivation is much more difficult to determine. To be amenable to study, the topic must be one that can be reasonably covered in a relatively short and focused questionnaire, with questions that are relevant for everyone in the sample. Self-administered questionnaires generally work best when the focus is in the present, rather than the past or the future. Unsupervised self-administered questionnaires, in particular, cannot be used in exploratory

studies or when the surveyor is in the process of developing a research question and the procedures by which that question will be studied.

It is particularly important for self-administered questionnaires to be user friendly. This is maximized both by the way questions are written and by the way in which questionnaires are formatted. Questions must be short and specific rather than general. This may mean that multiple questions, rather than a single question, must be asked. Response categories must be exhaustive and mutually exclusive. Sometimes, this means that questions should be structured so that respondents can select more than one answer to describe themselves, or that a “residual other” must be included among the response categories for a respondent. “Residual other” categories are particularly important when researchers do not think that they have an exhaustive list of possible answer categories.

One of the biggest errors made in developing questionnaires is to try to make them look shorter with the idea that this will increase response rates and decrease mailing costs. This is often done by squeezing all of the questions onto two pages, leaving little or no space between questions, and by using small type fonts. In fact, such procedures reduce response rates. Forcing questions onto a few pages generally means that the font is small and no space is left between questions. A small font makes it difficult for respondents to read the questionnaire. Judicious use of blank space makes it easier for respondents to see both the questions and the associated response categories, and to expeditiously progress through the questionnaire.

—Linda B. Bourque

REFERENCES

- Bourque, L. B., & Fielder, E. P. (2003). *How to conduct self-administered and mail surveys* (2nd ed.). Thousand Oaks, CA: Sage.
- Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 464–494.
- Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

SELF-COMPLETED QUESTIONNAIRE.

See SELF-ADMINISTERED QUESTIONNAIRE

SELF-REPORT MEASURE

Information about individuals may be obtained by direct OBSERVATION or through their reporting on their own behavior or attitudes in a survey. In some circumstances, the researcher may have a choice about which method of observation to use, but in many cases, self-report measures are the only ones that can be used. A researcher could ask individuals to get on a scale in order to measure their weight, or ask them in a SURVEY how much they weigh. In this case, the scale is much more likely to produce the correct value. Another researcher interested in voting behavior could check official records for voting information or could ask respondents in a survey whether or not they voted in the most recent election. Some voting studies use both measures in order to feel more confident about the VALIDITY of the respondents' actual behavior.

The potential problem with self-report measures is that they may contain one or more kinds of SYSTEMATIC ERROR or BIAS. This may derive from the way the question is worded or the response alternatives the respondent is offered, the order in which the questions were asked, the respondent's interaction with the INTERVIEWER, or the respondent's view of "acceptable" answers in this particular social setting.

OPEN-ENDED QUESTIONS, where respondents provide answers in their own words, often provide different responses than do CLOSED-ENDED QUESTIONS, where respondents have to choose from one of the offered responses. Offering an explicit opportunity to say "I don't know" or the alternative of saying "I haven't really thought about that" usually produces a different pattern of responses than not offering such an option. The range of scales and how they differ among questions can also explain differences in the answers respondents give (Schwarz, Hippler, Deutsch, & Strack, 1985).

The order in which questions are asked may lead respondents to an inevitable conclusion by the final item in the series. Asking a series of questions in an Agree/Disagree format may induce RESPONSE SET bias because of the willingness of respondents to agree with attitudinal statements posed to them (Schuman & Presser, 1996).

One of the most important forms of self-report bias is associated with SOCIAL DESIRABILITY BIAS, or the

tendency to offer responses that are felt to be more acceptable than the others. This is the most common explanation for the universally observed propensity of people to say they voted when record checks indicate that many of them did not. Racist attitudes are typically understated for the same reason. These self-report response patterns can be linked to the nature of the relationship between the social characteristics of the respondent and the interviewer, such as race or gender.

—Michael W. Traugott

REFERENCES

- Schuman, H., & Presser, S. (1996). *Questions & answers in attitude surveys*. Thousand Oaks, CA: Sage.
- Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, 49(3), 388–395.
- Traugott, M. W., & Katosh, J. P. (1979). Response validity in surveys of voting behavior. *Public Opinion Quarterly*, 43(3), 359–377.

SEMANTIC DIFFERENTIAL SCALE

The semantic differential (SD) is a technique developed during the 1940s and 1950s by Charles E. Osgood to MEASURE the meaning of language quantitatively. Words may have different meanings to different individuals as a function of their experiences in the world. For example, "poverty" has been experienced differently by 7- and 70-year-olds, and by the rich and the homeless. Their expressions of understanding of poverty are modified by these experiences. The SD captures these different meanings by providing some precision in how our understanding of words differs.

The typical semantic differential test requires a subject to assess a stimulus word in terms of a series of descriptive bipolar (e.g., good-bad) scales. The subject is asked to rate the stimulus between the extreme and opposing adjectives that define the ends of these scales. Typically, these bipolar scales have 5 or 7 points. The odd number allows the subject to choose a midpoint or neutral point (which would not be available if an even number of options were used). The various scale positions may be unlabeled, numbered, or labeled (see Table 1).

Table 1

Good	[]	[]	[]	[]	[]	Bad
Good	-2	-1	0	1	2	Bad
Good	Very	Quite	Neither	Quite	Very	Bad

Respondents seem to prefer the labeled version.

Typically, a stimulus will be assessed in terms of a large number of these bipolar scales. For example,

Good-Bad	Hard-Soft
Rough-Smooth	Fast-Slow
Fresh-Stale	Noisy-Quiet
Small-Large	Young-Old
Strong-Weak	Tense-Relaxed
Active-Passive	Valuable-Worthless
Cold-Hot	Unfair-Fair
Honest-Dishonest	Empty-Full
Bitter-Sweet	Cruel-Kind

The pattern of responses to these bipolar scales defines the meaning subjects attribute to the stimulus word.

Osgood's early research into the usefulness of the SD suggests that respondents' ratings of stimuli on these BIPOLAR SCALES tend to be correlated. Osgood argues that the semantic space is MULTIDIMENSIONAL but can be represented efficiently by a limited number of orthogonal DIMENSIONS. Three dimensions typically account for much of the covariation found in responses to the individual bipolar scales. These dimensions have been labeled Evaluation, Potency, and Activity (EPA) and have been found in a wide array of diverse studies. One way to interpret this finding is that, although the meanings of the bipolar scales may vary, many are essentially equivalent and hence may be represented by a small number of dimensions. To the extent that meanings are equivalent across groups, space, or time, the similarities and differences of groups, cultures, and epochs may be examined.

In practice, the semantic differential may be used to develop a nuanced understanding of attitudes. For example, Americans often hold ambiguous opinions of their national institutions and political leaders. Yet citizens are often asked their assessment of a political institution or elected official in a single question. One common question used to examine public perceptions of the American president is, "Do you

approve or disapprove of the way [name of current president] is handling his job as president?" Using the SD approach, it is possible to develop a more complex appraisal of public opinion by asking respondents to rate the president on a series of bipolar scales (see Table 2):

Table 2

Please rate George W. Bush on each of the following distinctions:

Good	[]	[]	[]	[]	[]	Bad
Weak	[]	[]	[]	[]	[]	Strong
Active	[]	[]	[]	[]	[]	Passive
Fair	[]	[]	[]	[]	[]	Unfair
Helpful	[]	[]	[]	[]	[]	Unhelpful
Dishonest	[]	[]	[]	[]	[]	Honest
Sharp	[]	[]	[]	[]	[]	Dull
Ineffective	[]	[]	[]	[]	[]	Effective
Fast	[]	[]	[]	[]	[]	Slow
Valuable	[]	[]	[]	[]	[]	Worthless
Powerless	[]	[]	[]	[]	[]	Powerful
Brave	[]	[]	[]	[]	[]	Cowardly

Responses to each scale may be examined individually, or they may be subject to data reduction (e.g., FACTOR ANALYSIS) to explore public assessments of the president on the EPA dimensions.

The SD has found wide use in sociology, communication studies, cognitive psychology, and marketing research.

Critics have identified one logical pitfall in the SD approach to measuring attitudes: Research based on the semantic differential begins with the assumption that people hold different attitudes/perceptions of a stimulus item—a word. But the SD technique assumes that adjectives chosen to represent the ends of the bipolar scales mean the same thing to each rater.

—John P. McIver

REFERENCES

- Heise, D. (1971). The semantic differential and attitude research. In G. F. Summers (Ed.), *Attitude measurement* (pp. 235–253). Chicago: Rand McNally.
- Osgood, C. E. (1952). The nature of and measurement of meaning. *Psychological Bulletin*, 49, 197–237.
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

- Osgood, C. E., & Tzeng, O. C. S. (Eds.). (1990). *Language, meaning, and culture: The selected papers of C. E. Osgood*. New York: Praeger.
- Snider, J. G., & Osgood, C. E. (Eds.). (1969). *Semantic differential technique: A sourcebook*. Chicago: Aldine.

SEMILOGARITHMIC. See LOGARITHMS

SEMIOTICS

Broadly defined, *semiotics* is the theory of meaning, the scientific domain that studies the structure and function of systems of signification and communication, and of the signifying (meaningful) wholes constructed from them. The term *semiotics* derives from *semi-otic*, used by one of the two founders of modern semiotics, the American philosopher Charles Sanders Peirce (1839–1914). A competing term is *semiology*, coined by the other founder of semiotics, the Swiss linguist Ferdinand de Saussure (1857–1913). *Semiotics* came to be used universally after a meeting held at Indiana University, Bloomington, in 1964 (although certain semioticians within the Saussurian tradition continue to differentiate between the two terms).

According to this definition, semiotics is not a separate science, but a scientific domain. In fact, the object of semiotics, signifying wholes and systems, is too broad to define a science; indeed, it cuts through a multiplicity of established sciences (e.g., linguistics, anthropology) and disciplines (e.g., literature, architecture, music).

For Saussurian STRUCTURAL LINGUISTICS, Peircian semiotics, and most other semiotic theories, the basic unit of semiotics is the unitary *sign* (although there are some theories that operate with broader signifying entities). A unitary sign is anything whatsoever that can be taken as an entity standing for (signifying) another entity. The word “tree” is a unitary sign, denoting a certain kind of plant (with a trunk, branches, leaves, etc.). Thus, unitary signs are signifying units, consisting in Saussure’s terminology of a *signifier*—the vehicle of signification—and a *signified*—the signification itself. The fact that semiotics studies the relationships between (at least) two planes—the signifier and the signified—differentiates semiotics from *information theory*, which deals only with the plane

of the signifier. Peirce defines the sign in terms of a triadic relational structure involving the *sign*, which is anything that conveys meaning; the *object* for which the sign stands; and the *interpretant*, the idea provoked by the object. The interpretant can itself be named by another sign, bringing about its own interpretant, and so on (theoretically) ad infinitum in a process of “unlimited semiosis.” The most fundamental division of signs for Peirce is into *icons*, *indices*, and *symbols*. Whereas the index and the icon are related to the things they signify pragmatically or through similarity, respectively, the symbol is purely conventional. In this, it resembles the Saussurian sign, which is conventional and arbitrary, that is, there is no natural relation between signs and the things for which they stand, a principle that also holds for the relation between signifier and signified.

Signs in communication are combined into signifying wholes, such as messages, discourses, or TEXTS. A signifying whole is a group of signs, combined according to certain syntactic rules. A building is an example of a signifying whole; a certain kind of building (with pointed arches, towers, and a set of other characteristics) denotes “gothic cathedral,” which in turn may call up connotations of “sacred place” and “transcendence.” This example shows the two levels of signification of a sign: the literal signification, or *denotation*, and the indirect or symbolic signification, *connotation*.

Signifying wholes presuppose systems of signification. Saussure called the system of signification *langue* (usually translated as “the language system” and used in semiotics to refer to any system of signs). The system is *synchronic*, that is, it represents a particular static state of *langue* (as opposed to its *diachronic* changes); it is social in nature, and, for Saussure, it is the privileged point of view for the study of natural language. The term *structure* is often used as equivalent to *system*, although it can also refer to the rules governing the system.

The system is a key concept for Saussurian semiotics. Given the principle of the arbitrariness of the sign, meaning is not considered as deriving from any relationship between signs and the real world, but is the result of the mutual relations between signs within the context of an abstract and purely formal system of an algebraic nature. Signifiers (and signifieds) are constituted through their differences from other signifiers (or signifieds) of the system. The differential nature of signification is one of the founding principles of Saussurian structural linguistics. Ultimately, any

positive content of the signified is denied, because signification follows from the relations of a sign with the other signs of the system, its *value*. Values—and thus the whole system—are determined through the interplay of two kinds of relations between signifiers: *paradigmatic* relations, deriving from the classification of signs in sets (one of Saussure's examples is the set “teaching,” “instruction,” “apprenticeship,” “education”), and *syntagmatic* relations, the result of selecting signs from various paradigms and combining them into the linear syntagms of language (sentences, texts).

Langue is opposed to *parole* (speaking), the use of the language system by individuals. Semiotic systems are used and thus function within society. Their central, although not their only, function is communication. Signifying wholes are constituted on the level of *parole*. *Pragmatics* (in the wide sense) is the branch of semiotics that studies the process of production and functioning of signifying wholes, that is, their use; their cognitive, emotional, or behavioral-corporeal impact on the receiver; and their dynamic interaction with the situation of communication.

HISTORICAL DEVELOPMENT

Although the Greeks and other ancient cultures were concerned with the issue of meaning, modern semiotics has its origins in the works of Peirce and Saussure. In the United States, semiotics gained wider recognition through the work of Charles William Morris, who divided semiotic analysis into *syntactics*, *semantics*, and *pragmatics* (which he defines as the study of the relation between signs and interlocutors). A tendency primarily in the United States to extend semiotics beyond the study of human signifying systems to animal communication (*zoosemiotics*) and further to the natural processes of all living organisms (*biosemiotics*) owes much to the work of Thomas A. Sebeok.

A decisive influence on the formation of European semiotics was the Russian linguist Roman Jakobson. In the second decade of the 20th century, Jakobson, who worked within the Saussurian tradition, participated in the creation of two semiotic groups in Russia that formed the nucleus of Russian Formalism, and in the next decade, he was instrumental in founding the Prague Linguistic Circle, which marked the constitution of European linguistic STRUCTURALISM proper. The work of Jakobson and the Prague Circle envisages natural language as one among a multitude of sign

systems, which together constitute culture as a complex system of communication (a view later continued by the postwar Moscow-Tartu School of semiotics). An important contribution to the development of European semiotics was also made by the Danish linguist Louis Hjelmslev, who cofounded the Linguistic Circle of Copenhagen in 1931.

During World War II, Jakobson met and influenced a French anthropologist, Claude Lévi-Strauss. The structural anthropology of Lévi-Strauss was crucial to the flourishing of semiotics in the 1960s and 1970s. In 1964, the literary scholar Roland Barthes published the first contemporary treatise on semiotics, *Éléments de Sémiologie*, soon followed by major works by other well-known semioticians such as the founder of the Paris School Algirdas Julien Greimas and the Italian philosopher Umberto Eco. At the same time, the tendency that later became known as POSTSTRUCTURALISM developed in Paris; eminent authors are the philosophers Jacques Derrida and François Lyotard, the psychoanalyst Jacques Lacan, the historian of knowledge (of “discourses”) Michel Foucault, and the sociologist Jean Baudrillard.

METHODOLOGY AND APPLICATIONS

Semiotic methodology can be understood by comparison to two of the methods most commonly used today for the analysis of meaning, CONTENT ANALYSIS and the SEMANTIC DIFFERENTIAL. Both of these presuppose the creation of qualitative categories, and both result in quantitative statistical analysis; however, they adopt a rather limited view of the text and lack any substantial theoretical elaboration, drawing their value from their methodological, not to say technical, interest. Contrary to this, semiotic methodology is based on an extremely sophisticated theory, covers all or nearly all of the aspects of a text, and has been established as primarily qualitative.

Conventional methods for the analysis of meaning do not generally address the syntagmatic, or linear, dimension of texts. On the other hand, semiotics disposes of a quite elaborate *narrative* theory formulated largely by A. J. Greimas. Greimas decomposes a text into units of “doing,” integrating units of “state” formed by the conjunction or disjunction between various kinds of subject and object. This elementary narrative program, the act, produces or changes a state, that is, effects the transition from one state to another. The syntagmatic organization of these acts

leads to action. Every discourse—not only narrative discourse—is a process leading from an initial to a final state and can be formally simulated by an ALGORITHM of transformation.

One of the earliest examples of semiotic methodology is offered by Barthes's (1977) paradigmatic analysis of an advertisement by the Italian food manufacturer Panzani. The aim of paradigmatic analysis is comparable to that of content analysis and the semantic differential.

Barthes distinguishes two complementary types of messages in the advertisement, the iconic and the linguistic. In the iconic message, he identifies two levels, the literal or denotative message (i.e., the objects in the image), and the symbolic or connotative message; the linguistic message serves to guide the interpretation of the symbolic. Barthes observes that the literal message as syntagm is the necessary support of the symbolic one, because certain noncontinuous elements of the former lead to the appearance of "strong" signs, bearers of a second-level meaning. He identifies four major meanings of this kind: "return from the market" with the accompanying euphoric values of the freshness of the produce and their domestic use, anchored in the half-open net bag from which the provisions spill; "italicity" (cf. the name "Panzani"), a kind of condensed essence of everything Italian, from spaghetti to painting, derived from the three colors in the image together with the (Mediterranean) vegetables tomato and pepper (green, white and red being the colors of the Italian flag); "complete culinary service," due to the gathering together of many different foodstuffs; and "still life," due to the composition, which is reminiscent of painting. To these four symbolic meanings correspond four paradigmatic codes, which Barthes calls the practical, the national, the cultural, and the aesthetic.

Although semiotic analysis is usually qualitative, it is occasionally combined with a quantitative aspect. A. Ph. Lagopoulos and K. Boklund-Lagopoulou, studying the conception of regional space in four settlements of northern Greece, analyze the content of the interviews they conducted into 32 paradigmatic codes, grouped into eight major sets, characteristic of "discourse on space." They then use frequency counts and cross-tabulations to study a wide set of semiotic (conceptual) variables, including the codes, and their relationship to sociological (social class, gender, age group, education) and sociogeographical (urban vs. rural, settlement hierarchy) variables, as well as relationships between various kinds of semiotic data.

SEMIOTICS TODAY

Today, semiotics has permeated all of the social and human sciences and is an important component of the dominant paradigm of our times, POSTMODERNISM. In linguistics and literary studies, it has contributed to the development of such areas as lexicology; DISCOURSE ANALYSIS (in particular, rhetoric); narratology; and stylistics. It has deeply influenced anthropology and the study of mythology. It has also penetrated semantics and contributed greatly to the study of nonverbal and mixed systems, leading to the creation of, for example, a semiotics of behavior, painting and photography, music, architecture and urban space, culinary habits, the circus, fashion, advertising, cinema, and the theater. Most significantly, semiotics has provided a unified theoretical framework for the comparative study of cultural phenomena, contributing significantly to the present unparalleled flourishing of interdisciplinary and cross-disciplinary studies of signifying structures and practices of all kinds.

—Karin Boklund-Lagopoulos and
Alexandros Lagopoulos

REFERENCES

- Barthes, R. (1967). *Elements of semiology*. London: Jonathan Cape.
- Barthes, R. (1977). Rhetoric of the image. In *Image, music, text* (S. Heath, Trans.). London: Fontana.
- de Saussure, F. (1979). *Cours de linguistique générale* (T. de Mauro, Ed.). Paris: Payot.
- Eco, U. (1976). *A theory of semiotics*. Bloomington: Indiana University Press.
- Greimas, A. J., & Courtés, J. (1982). *Semiotics and language: An analytical dictionary*. Bloomington: Indiana University Press.
- Lagopoulos, A. Ph., & Boklund-Lagopoulos, K. (1992). *Meaning and geography: The social conception of the region in northern Greece*. Berlin: Mouton de Gruyter.
- Peirce, C. S. (1931–1958). *Collected papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press.

SEMIPARTIAL CORRELATION

A semipartial correlation is a statistic measuring the relationship between a DEPENDENT VARIABLE (or criterion) with an adjusted INDEPENDENT VARIABLE (or predictor) controlling for the effect of other predictors in the analysis. Like PARTIAL CORRELATION, the

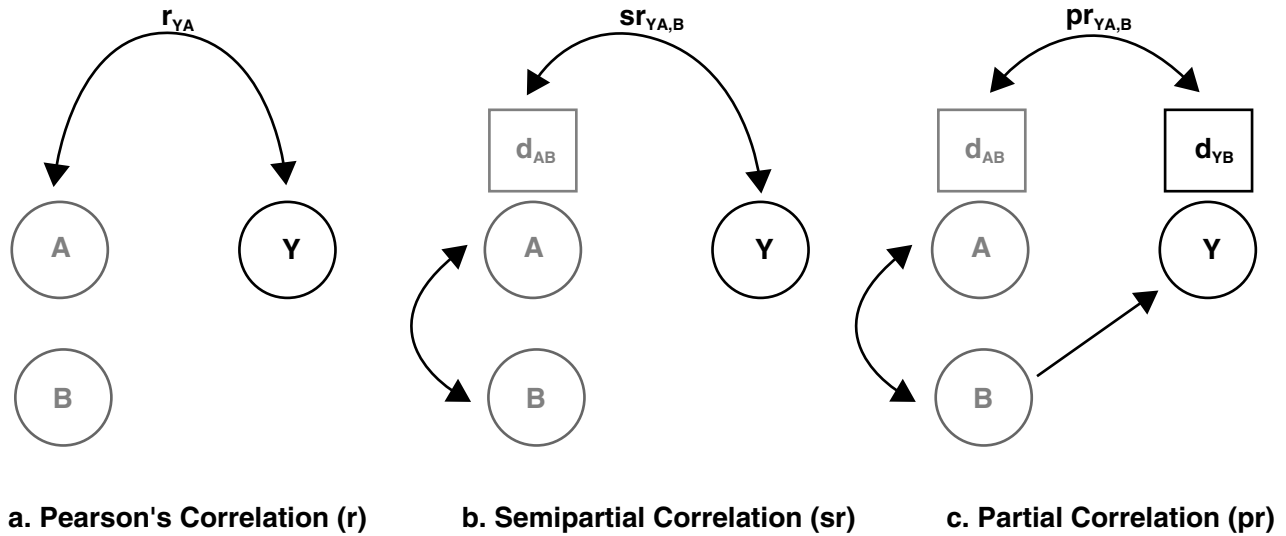


Figure 1 Diagrams Illustrating Conceptual Differences Between Semipartial, Partial, and Pearson's Correlations

semipartial correlation is computed in the process of MULTIPLE REGRESSION ANALYSIS.

The diagrams illustrate how the semipartial correlation differs conceptually from both partial and Pearson's correlations. Grey circles (A s and B s) in the diagrams represent predictor variables, whereas the black circles, Y s, represent criterion variables. Boxes represent residuals of A s and Y s after subtracting the effects that the B s exert on them. Curved arrows represent correlation (or COVARIANCE) between pairs of variables in the diagrams, whereas the straight arrow represents a regression path connecting B to Y .

In Figure 1a, the curved arrow r_{YA} represents a zero-order Pearson's correlation coefficient, which measures the extent to which A and Y are correlated with each other, not taking into account the effect of the third variable B . In Figure 1b, the curved arrow $sr_{YA,B}$ represents a semipartial correlation coefficient. It measures the extent to which the criterion, Y , and the RESIDUAL of A (i.e., $d_{A,B}$), obtained by subtracting from A its covariance with B , are correlated with each other. In Figure 1c, the curved arrow $pr_{YA,B}$ represents a partial correlation coefficient. It measures the extent to which the residual of Y (i.e., $d_{Y,B}$) and the residual of A (i.e., $d_{A,B}$) are correlated with each other.

Illustratively, in a multiple regression analysis in which Y is the criterion variable that is predicted by A and B (two predictors), both the semipartial correlation and the partial correlation obtained from A and B would be related to specific (or unique) contributions

of A and B , respectively. In particular, the squared semipartial correlation for A is the proportion of the *total* variance of Y uniquely accounted for by A , after removing the shared variance between A and B from A but not from Y . Alternatively, the squared partial correlation for A is the proportion of the *residual* variance of Y uniquely accounted for by A after removing the effect of B on both A and Y .

Both the partial and semipartial correlations for any given pair of variables can be calculated on the basis of the zero-order Pearson's correlation between a pair of variables according to the following algebraic rules:

$$sr_{YA,B} = \frac{r_{AY} - (r_{AB})(r_{YB})}{\sqrt{1 - r_{AB}^2}} \quad (1)$$

$$pr_{YA,B} = \frac{r_{AY} - (r_{AB})(r_{YB})}{\sqrt{1 - r_{YB}^2} \times \sqrt{1 - r_{AB}^2}} \quad (2)$$

Comparing equations (1) and (2), the computation of a semipartial correlation (i.e., $sr_{YA,B}$) differs from that of a partial correlation (i.e., $pr_{YA,B}$) in the composition of their denominators. The partial correlation equation has a smaller denominator than does the semipartial correlation equation (i.e., for the partial correlation, the denominator is a product of two terms each smaller than 1.00, whereas the semipartial correlation includes only one of those terms). These differences

make a semipartial correlation always lower than a partial correlation, except for the cases where *A* and *B* are independent or uncorrelated.

—Marco Lauriola

REFERENCES

- Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: Holt, Rinehart and Winston.

SEMISTRUCTURED INTERVIEW

Semistructured interviewing is an overarching term used to describe a range of different forms of interviewing most commonly associated with QUALITATIVE RESEARCH. The defining characteristic of semistructured interviews is that they have a flexible and fluid structure, unlike STRUCTURED INTERVIEWS, which contain a structured sequence of questions to be asked in the same way of all interviewees. The structure of a semistructured interview is usually organized around an aide memoire or INTERVIEW GUIDE. This contains topics, themes, or areas to be covered during the course of the interview, rather than a sequenced script of standardized questions. The aim is usually to ensure flexibility in how and in what sequence questions are asked, and in whether and how particular areas might be followed up and developed with different interviewees. This is so that the interview can be shaped by the interviewee's own understandings as well as the researcher's interests, and unexpected themes can emerge.

LOGIC AND UNDERPINNINGS OF SEMISTRUCTURED INTERVIEWING

Semistructured interviewing is most closely associated with INTERPRETIVIST, INTERACTIONIST, CONSTRUCTIONIST, FEMINIST, PSYCHOANALYTIC, and ORAL or LIFE HISTORY traditions in the social sciences. It reflects an ONTOLOGICAL position that is concerned with people's knowledge, understandings, interpretations, experiences, and interactions. Elements of the different traditions can be identified in contemporary semistructured interviewing practice, which Jennifer Mason suggests is broadly characterized by

the interactional exchange of dialogue; a relatively informal style; a thematic, topic-centered, biographical, or narrative approach; and the belief that knowledge is situated and contextual, and that therefore the role of the interview is to ensure that relevant contexts are brought into focus so that situated knowledge can be produced (Mason, 2002, p. 62).

Most commentators agree that the logic of semistructured interviewing is to generate data interactively, and Steinar Kvale has described qualitative research interviews as "a construction site of knowledge" (Kvale, 1996, p. 2). This means that the interviewer, and not just the interviewee, is seen to have an active, REFLEXIVE, and constitutive role in the process of knowledge construction. Data are thus seen to be derived from the interaction between interviewer and interviewee, and it is that interaction, rather than simply the "answers" given by the interviewee, that should be analyzed.

The relatively open, flexible, and interactive approach to interview structure is generally intended to generate interviewees' accounts of their own perspectives, perceptions, experiences, understandings, interpretations, and interactions. A more standardized and structured approach might overly impose the researcher's own framework of meaning and understanding onto the consequent data. It might also risk overlooking events and experiences that are important from the interviewees' point of view, that are relevant to the research but have not been anticipated, or that are particular to interviewees' own biographies or ways of perceiving. However, some form of aide memoire or interview guide is usually used by the interviewer to ensure that the interview addresses themes identified in advance as relevant to the research.

Semistructured interviewing sometimes implies a particular approach to research ETHICS, whereby the power relationship between interviewer and interviewee is equalized as much as possible, and where the interviewee gets plenty of opportunity to tell his or her story in his or her own way. Although this might make semistructured interviewing appear to be more ethical than standardized and structured methods, it is also recognized that the data generated can be very personal and sensitive, especially where they have emerged in a relationship of trust and close RAPPORT between interviewer and interviewee. Therefore, questions of confidentiality and ethics are not straightforward in this context.

TYPES OF SEMISTRUCTURED INTERVIEW

In practice, there is a range of types of semistructured interviewing. Differences between types tend to center around the extent to which the interviewer is seen to structure and guide the interaction; the nature of the organizing logic for the interview, such as a life story or biography, or a loose topic or set of topics; the extent to which the consequent data are seen to be and are analyzed as interactions or as interviewees' accounts; and the extent to which the interview data speak for themselves or need to be interpreted, by whom and through what mechanisms. Three distinct types of semistructured interview are the following:

1. *The ethnographic interview* is usually part of a broader piece of field research. Ethnographic interviews are often conducted within ongoing relationships in the field, and data of a variety of forms are thus generated both within and outside the interview. They can involve sustained contact between interviewer and interviewee over a period of time, thus extending the possibilities for interactive coproduction of knowledge and for rapport between the parties. In analyzing the consequent data, there is strong emphasis on the interviewees' own interpretations and understandings.

2. *The psychoanalytic interview* is sometimes, although not of necessity, conceived of as part of a therapeutic encounter. The aim is usually to explore not just the interviewees' conscious but also their subconscious understandings. Wendy Hollway and Tony Jefferson (2000) recommend the use of the free association narrative method, which uses a minimum of interviewer-imposed structure to facilitate the researcher's understanding of subconscious intersubjectivity and the psychosocial subject through analysis of interviewees' patterns of free association. The interviewees' own interpretations are thus not taken at face value, and the researcher deploys psychoanalytic theory to analyze the data.

3. *The life or oral history interview* involves generating biographies and life stories. Life histories can be generated in a range of ways, including documentary or visual methods such as analysis of diaries, letters, and photographs. These can become part of the interview, so that, for example, interviewees may be asked to talk about their family photographs. Some approaches use these methods to document lives as they are told, in a fairly noninterpretive way. Others emphasize interpretive and interactive elements in the construction of life

history knowledge. As with ethnographic interviews, life history interviews are sometimes conducted within ongoing relationships over a period of time.

EVALUATION OF SEMISTRUCTURED INTERVIEWING

Semistructured interviewing has been demonstrated to be a valuable method for gaining interpretive data. Criticisms that it produces data that cannot permit comparisons between interviews because they are not standardized are misplaced, because it uses a logic where comparison is based on the fullness of understanding of each case, rather than standardization of the data across cases. However, semistructured interviewing alone can produce only partial interpretive understandings and can be usefully supplemented by other methods, such as those that can extend the situational dimensions of knowledge, including PARTICIPANT OBSERVATION and VISUAL METHODS.

—Jennifer Mason

REFERENCES

- Hollway, W., & Jefferson, T. (2000). *Doing qualitative research differently*. London: Sage.
- Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. London: Sage.
- Mason, J. (2002). *Qualitative researching* (2nd ed.). London: Sage.
- Plummer, K. (2001). *Documents of life 2: An invitation to a critical humanism*. London: Sage.

SENSITIVE TOPICS, RESEARCHING

Researchers, not uncommonly, must study topics that are sensitive in some way. There is no definitive list of sensitive topics. Often, sensitivity seems to emerge out of the relationship between the topic under consideration and the social context within which that topic is studied. Seemingly difficult topics can be studied without problems, whereas apparently innocuous topics prove to be highly sensitive. Thus, groups that might seem at first glance to be strange, bizarre, or even frightening sometimes feel misunderstood by the wider society, and regard with equanimity the presence of a researcher whom they feel might convey their worldview to a wider audience. On the other hand,

some apparently normal settings can conceal patterns of deviance participants wish to shield from scrutiny. In these situations, ACCESS and FIELD RELATIONS can become unexpectedly problematic.

Sensitive topics often involve a level of threat or risk to those studied that renders problematic the collection, holding, and/or dissemination of research data. Often, such threats arise from a researcher's interest in particular areas of social life. Research may be sensitive where it touches on aspects of life such as sexual activity, bereavement, or intensely held personal beliefs that are private, stressful, or sacred. The possibility in the eyes of those studied that research might reveal information about research participants that is stigmatizing or incriminating in some way is frequently a source of sensitivity. Research is often sensitive, too, where it impinges on the vested interests of powerful people or institutions, or the exercise of coercion or domination. Research can also pose threats to the researcher. Researchers who study deviant groups sometimes come to share the social stigma attached to those being studied. Friends, colleagues, and university administrators sometimes assume, for example, that researchers investigating aspects of sexual behavior share the predilections of those they study. Physical danger can arise in situations of violent social conflict or when researching groups within which violence is common.

Because they often involve contingencies that are less common in other kinds of study, sensitive topics pose difficult methodological and technical problems. Access to field settings is frequently problematic. Those who control entry to a setting will often decline to provide assistance to the researcher, or surround access with conditions or stipulations. In sensitive contexts, the relationship between the researcher and the researched can become hedged about with mistrust, concealment, and dissimulation in ways that affect the reliability and validity of data. In these situations, acceptance of the researcher is often dependent on the lengthy inculcation of trustful relations with research participants. It is possible to find technical or procedural solutions to the problems caused by the threatening character of the research process, such as collecting data under strongly anonymous conditions. Despite the serious ethical problems involved, some researchers have adopted robust investigative tactics, including covert methods or forceful interviewing strategies, to overcome the tendency of powerful groups to conceal their activities from outside scrutiny.

Sensitive topics raise sometimes difficult issues related to the ethics, politics, and legal aspects of research. Those researching sensitive topics may need to be more acutely aware of their ethical responsibilities to research participants than would be the case with the study of a more innocuous topic. Special care might need to be taken, for example, with anonymization and data protection. Increasingly, research participants have enhanced legal rights vis-à-vis researchers. A number of countries now have regulatory systems that mandate prior ethical review of social research. Sometimes, though, regulatory frameworks hinder research on sensitive topics in ways that arguably run counter to the spirit in which they were framed. Studies deemed too sensitive have sometimes run into difficulty with regulators. Researchers now often enter collaborative research relationships that allow powerless or marginalized groups to voice their concerns and aspirations. PARTICIPATORY or ACTION RESEARCH methods allow for the development of egalitarian and reciprocal relationships between researchers and research participants. Such relationships, however, are not immune to tensions between researcher and researched, or pressures toward co-option.

Sensitive topics tax the ingenuity of researchers and can make substantial personal demands on them. Skill, tenacity, and imagination are all required to cope successfully with the problems and issues that sensitive topics raise. Although these have sometimes encouraged researchers to innovate in areas such as asking sensitive questions on surveys and protecting sensitive data, the relationship between methodological or technical decisions in research and ethical, legal, or political considerations is often complex. Although some researchers deal with the problematic aspects of studying sensitive topics by opting out of such areas altogether, researching sensitive topics remains a challenging and necessary activity for social research.

—Raymond M. Lee

See also ACTION RESEARCH, ANONYMITY, COVERT RESEARCH, DANGER IN RESEARCH, FIELD RELATIONS, POLITICS OF RESEARCH

REFERENCE

- Lee, R. M. (1993). *Doing research on sensitive topics*. London: Sage.

SENSITIZING CONCEPT

The American sociologist Herbert Blumer, founder of SYMBOLIC INTERACTIONISM, contrasted sensitizing concepts with the definitive concepts insisted on by many QUANTITATIVE social researchers. Sensitizing concepts do not involve using “fixed and specific procedures” to identify a set of phenomena, but instead give “a general sense of reference and guidance in approaching empirical instances.” So, whereas definitive concepts “provide prescriptions of what to see, sensitizing concepts merely suggest directions along which to look” (Blumer, 1969, p.148).

Blumer recognized that vagueness of theoretical concepts can be a problem. However, he denied that the solution was developing definitive concepts, as suggested by the POSITIVIST philosophy of science that had become influential within American sociology in the 1930s, 1940s, and 1950s (see OPERATIONISM). His championing of sensitizing concepts reflected the PRAGMATIST philosophy from which his own thinking about methodology grew. Pragmatists, notably Dewey, argued that concepts are tools for use and should be judged according to their instrumental value in enabling people to understand, and to act on, the world. Also crucial to Blumer’s argument is the idea that concepts have to be developed and refined over the course of research if they are to capture the nature of the phenomena to which they relate. He argued that beginning research by stating precise definitions of concepts, as advocated by many positivists, amounts to a spurious scientificity and is likely to obstruct the task of gaining genuine knowledge about the social world.

At the same time, there are suggestions within Blumer’s work of the more radical view that vagueness, in some sense, may be an inevitable feature of social scientific concepts. One reason for this is the difficulty of capturing the diverse, processual, and contingent nature of social phenomena. Blumer argues that this seems to imply inevitable reliance on tacit cultural knowledge, on *VERSTEHEN*. Thus, he writes that only physical objects can be “translated into a space-time framework or brought inside of what George Mead has called the touch-sight field,” whereas the observation of social phenomena requires “a judgement based on sensing the social relations of the situation in which the behavior occurs and on applying some social norm present in the experience of the observer” (Blumer, 1969, p. 178).

Blumer recognizes that most of the key concepts of sociology fulfill an important sensitizing role. However, he insists that they need to be clarified, as far as this is possible, through researchers enriching their firsthand experience of relevant areas in the social world. This is to be done, above all, by what he calls NATURALISTIC RESEARCH—qualitative or ethnographic inquiry—in which empirical instances are subjected to “careful flexible scrutiny” in their natural context, “with an eye to disengaging the generic nature” of the concept (Blumer, 1969, p. 44).

—Martyn Hammersley

REFERENCES

- Blumer, H. (1969). *Symbolic interactionism*. Englewood Cliffs, NJ: Prentice Hall.
- Hammersley, M. (1989). The problem of the concept: Herbert Blumer on the relationship between concepts and data. *Journal of Contemporary Ethnography*, 18, 133–160.

SENTENCE COMPLETION TEST

A sentence completion test entails filling in one or more missing words in a sentence. A simple example is the following:

She poured him a cup of tea and offered a slice of ____.

The ability to choose semantically appropriate completions is taken to indicate degree of linguistic ability. Alternative versions of the test provide two missing words in order to test understanding of causal or logical relationships, or a full paragraph with multiple missing words.

Sentence completion, or “cloze,” tests are widely used in educational testing (e.g., as part of the American SAT and Graduate Record Exams, and of the Test of English as a Foreign Language). In testing, students are given a range of possible word completions. All are syntactically sound, but only one is semantically appropriate. The validity of completion as a measure of linguistic ability depends on the assumption that understanding the meaning of one term in a natural language requires a supporting structure of many other semantically related terms. Consequently, subjects cannot train to the test because reliable completion performance depends on the degree to which the whole structure is already in place.

In psychological research, subjects are seldom offered multiple completions, but rather allowed to complete the sentence as they think appropriate. The DEPENDENT VARIABLE of interest is then their choice of word, or the time taken to complete the sentence.

Psychological applications divide into two types, depending on whether completions are used as an indirect measure of a nonlinguistic psychological state (e.g., an aspect of personality) or as a tool for investigating language processing itself (e.g., in testing theories of sentence comprehension). In nonlinguistic work, completions provide information about a subject's psychological state only when compared to previously collected completion norms that have been validated for measuring that state.

In psycholinguistics, sentence completion is used in conjunction with free association and priming tasks to quantify otherwise inaccessible and unconscious linguistic expectations (Garman, 1990). In an association task, the subject is given a single word cue and must generate all the words that come to mind. The probability that each word in a language will appear among the first several words after a cue can be estimated from a large body of text. Subjects tend to generate words whose conditional probability given the cue is high. Priming refers to the reduction in time taken to recognize a word caused by brief, prior presentation of a semantically related word, compared to a semantically unrelated one. Priming effects are well predicted by the level of substitutability between the word that is recognized and the prime word. Substitutability is a measure of how easily one word can be replaced with another in context without making the resulting sentences semantically incongruous, and it can be quantified by measuring distributional similarity, the extent to which words share verbal context (Redington & Chater, 1997). Substitutability accounts of meaning are inherently relational and formalize the testing assumption that understanding meaning presupposes substantial surrounding linguistic structure (Cruse, 1986).

Physically, unrelated words in the priming task and incongruous substitutions into sentential contexts trigger large N400 components in brain event-related potentials, lengthened gaze duration, and increased reaction times.

Because the association task is a compressed form of sentence completion, and priming results are predicted by a substitutability measure closely related to that tapped by sentence completion, it is likely that

sentence completion also depends fundamentally on statistical measures of verbal context. On this view, second language applications of the sentence completion task measure how well subjects have internalized the substitutability intuitions that are generated by the underlying statistical properties of the second language and are a necessary by-product of learning to speak it.

—Will Lowe

REFERENCES

- Cruse, D. A. (1986). *Lexical semantics*. Cambridge, UK: Cambridge University Press.
- Garman, M. (1990). *Psycholinguistics*. Cambridge, UK: Cambridge University Press.
- Redington, M., & Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in the Cognitive Sciences*, 1(7), 273–281.

SEQUENTIAL ANALYSIS

In the context of social science research, sequential analysis refers to ways of organizing and analyzing behavioral data in an attempt to reveal sequential patterns or regularities in the observed behavior. Thus, unlike LOG-LINEAR or MULTIPLE REGRESSION ANALYSIS, for example, sequential analysis is not a particular statistical technique as such but a particular application of existing statistical techniques.

Data for sequential analysis are usually derived from SYSTEMATIC OBSERVATION (Bakeman & Gottman, 1997), are categorical in nature, and can be quite voluminous. Many of the statistical techniques used are those appropriate for CATEGORICAL DATA ANALYSIS generally, such as chi-square or log-linear analysis, although often a productive strategy is to derive from the data indexes of sequential patterning, which can then be analyzed with any appropriate statistical procedure (Bakeman & Quera, 1995). Thus, a sequential analysis can provide basic descriptive information about sequences of observed behavior. At the same time, it can reduce such data to a few theoretically targeted indexes of sequential process that then can be subjected to any appropriate analytic technique.

HISTORICAL DEVELOPMENT

The development of sequential analysis was spurred in the 1960s by the influence of human and

animal ETHOLOGY and was led largely by researchers interested in animal behavior and human development. Consistent with ethological approaches generally, sequential analysis was viewed as an inductive enterprise: Observers would construct an ethogram (a list of behaviors an animal might perform) and then record the order in which behaviors occurred. Such data consisted of *event sequences*: If the ethogram (or coding scheme) consisted of five codes (e.g., A, B, C, D, E), a data sequence might be A B C B C D E C E C D A, and so on. The notion was, with sufficient data and the right analysis, patterns would emerge.

Thus, researchers formed cross-classified tables: Rows represented the given or current code (Lag 0), columns represented the target or next code (Lag 1), successive pairs of codes in the sequence were tallied (first c_1c_2 , then c_2c_3 , c_3c_4 , etc., where c_i represents the code in the i th position), and cells whose observed frequency exceeded their expected frequency were noted. The extent to which observed frequency exceeds expected frequency can be gauged with an adjusted z score, a common CONTINGENCY TABLE statistic. When codes are many and no targeted hypothesis limits the codes examined, this approach can become unwieldy. For example, in an influential paper from the pre-computer revolution mid-1960s, Stuart A. Altman described covering the floor with patched-together sheets of paper describing matrices with hundreds of rows and columns, down on his hands and knees, looking for patterns.

LAGGED EVENT SEQUENCES

Given event sequences, typically Lag 1 tables prove most useful, but Lag 2, Lag 3, and so on (columns represent second behavior, third behavior, etc., after current one), can also be considered in an arrangement that Gene P. Sackett (1979) termed lag-sequential analysis. For example, if B occurred significantly more often after A at Lag 1, and C occurred significantly more often after A at Lag 2, this would suggest that A B C was a common sequence, especially if C occurred significantly more often after B at Lag 1. Such analyses are most useful when the duration of events does not vary much and when order and not timing is of paramount concern.

TIMED SEQUENCES: AN EXAMPLE

The previous example assumed that events were coded in sequence but that their onsets or offsets were

Table 1 Toddler's Onsets of Rhythmic Vocalizations Cross-Classified by Whether or Not They Occurred Within 5 Seconds of Mother's Onsets

<i>Within 5 s of Mother's Onset</i>	<i>Toddler Onset</i>		<i>Totals</i>
	<i>Yes</i>	<i>No</i>	
Yes	11	189	200
No	29	97	1,000
Totals	40	1,160	1,200

NOTE: This mother-toddler dyad was observed for 1,200 seconds or 20 minutes; the tallying unit is the second.

not timed. When timing information is available—and the current digital revolution makes timing information ever easier to record—sequences can be represented and analyzed in ways that provide analytic opportunities not available with event sequences. Assume that times are recorded to the nearest second. Then, *timed sequences* can consist of any number of mutually exclusive or co-occurring behaviors, and the time unit, not the event, can be used as the tallying unit when constructing contingency tables. This can be very useful. Often, researchers want to know whether one behavior occurred within a specified time relative to another and are not particularly concerned with its lag position, that is, with whether other behaviors intervened.

For example, Deckner, Adamson, and Bakeman (2003) wanted to know whether mothers and their toddlers matched each other's rhythmic vocalizations and so coded onset and offset times for mothers' and toddlers' rhythmic vocalizations. Table 1 shows results for one dyad.

For the rows, seconds were classified as within a 5-second window defined by the second the mother began a rhythmic vocalization or not; for the columns, seconds were classified as a second the toddler began a rhythmic vocalization or not. A useful way to summarize this 2×2 table is to note that the odds the toddler began a rhythmic vocalization within 5 seconds of his or her mother beginning one were 0.0582 to 1 or 1 to 17.2, whereas the corresponding odds otherwise were 0.0299 to 1 or 1 to 33.5. Thus, the ODDS RATIO—a statistic probably used more in epidemiology than in other social science fields—is 1.95.

The ODDS RATIO is useful on two counts: First, it is useful descriptively to say how much greater the odds are that a behavior will begin within a specified time window than not. Second, the natural logarithm

of the odds ratio, which varies from minus to plus infinity with zero indicating no effect, is an excellent score for standard statistical analyses (the odds ratio itself, which varies from zero to plus infinity with 1 representing no effect, is not). Thus, Deckner et al. could report that 24-month-old females were more likely to match their mother's rhythmic vocalization than 24-month-old males or either male or female 18-month-old toddlers.

In sum, Deckner et al.'s study provides an excellent example of how sequential analysis can proceed with timed sequences. Onset and offset times for events are recoded; a computer program (Bakeman & Quera, 1995) then tallies seconds and computes indexes of sequential process (here, an odds ratio) for individual cases; and finally, a standard statistical technique (here, mixed model ANALYSIS OF VARIANCE) is applied to the sequential scores (here, the natural logarithm of the odds ratio).

—Roger Bakeman

REFERENCES

- Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.
- Bakeman, R., & Quera, V. (1995). *Analyzing interaction: Sequential analysis with SDIS and GSEQ*. New York: Cambridge University Press.
- Deckner, D. F., Adamson, L. B., & Bakeman, R. (2003). Rhythm in mother-infant interactions. *Infancy*, 4, 201–217.
- Sackett, G. P. (1979). The lag sequential analysis of contingency and cyclicity in behavioral interaction research. In J. D. Osofsky (Ed.), *Handbook of infant development* (1st ed., pp. 623–649). New York: Wiley.

SEQUENTIAL SAMPLING

Sequential SAMPLING involves drawing multiple SAMPLES and administering them one after another until, cumulatively, they meet the objectives of the survey. Frequently, it is difficult or impossible to know in advance what the response rate or cost per interview will be for a particular survey, especially for special POPULATIONS or when response patterns are changing. In these cases, it is difficult to accurately project the size of the sample that must be drawn in order to meet research objectives such as obtaining the desired number of completed responses or staying within the budget for data collection. Sequential sampling

relies on the principle that PROBABILITY SAMPLES drawn from the same study population can be combined and estimates calculated as if they were initially drawn as a single sample. Sequential sampling allows researchers to control more precisely the number of completed responses obtained by a survey without unduly compromising the criteria of probability selection or absence of NONRESPONSE (Kish, 1965).

Sequential sampling is often implemented by drawing three samples: an initial sample of one-half the size that could be optimistically expected to meet the study objectives; an additional sample of the same size as the first drawn from the same population; and a final, supplemental sample that is drawn from the same study population. The sample that is optimistically expected to meet the study objectives can be calculated using somewhat but not too ambitious response rates. This is generally done by dividing the response rate into the number of completed responses desired by the survey and using this as the optimistic sample size for data collection ($n_o = n/r$, where n_o is the optimistic size of the sample needed for data collection, n is the number of completed responses desired, and r is an optimistic estimate of the response rate). In practice, the first sample ($n_o/2$) is fully administered using all of the callback procedures specified in the survey design. When the data collection for the first sample is nearly completed, the second sample is placed in the field and data collection proceeds as planned. Researchers calculate estimates of nonresponse and cost per interview from the first sample. Using these estimates, the size of the third sample is calculated, the third sample is drawn from the same study population, and it is placed in the field for data collection according to the survey design. Sequential sampling allows researchers to hedge their bets against uncertain outcomes.

Although it provides obvious benefits, sequential sampling can increase costs of field operations and make survey administration more difficult. A common problem with sequential sampling has been the interruption of field operations while the results from the first sample were tabulated. The use of the second sample in the steps above minimizes the disruption to field operations. In addition, it is now common for interviewers to enter data directly into automated files, reducing the time needed for the calculations that are required to establish the size of the final sample.

Although the three-sample, sequential sample, described above, is generally adequate for obtaining CROSS-SECTIONAL survey data, the technique has

been expanded in two ways. First, telephone survey administrators often obtain multiple samples or replicates from commercial sampling firms that are administered as sequential samples, usually with several replicates being administered at the beginning of the survey and additional samples being activated for data collection as response rates are estimated from the first batch of samples. Because commercial sampling firms constantly update their files of useable telephone numbers and charge set-up fees for each separate sample, many telephone survey units routinely obtain excess samples in several replicates to maximize control over the uncertainties of data collection while controlling their costs.

Second, sequential sampling techniques are employed in tracking polls or rolling sample surveys that allow both cross-sectional and TIME-SERIES estimates from the study population. Tracking polls are used primarily in political campaigns by candidates, media outlets, and scholars (John, 1989; Moore, 1999; Rhee, 1996). Tracking polls allow estimates of the current level of support for a candidate and change in the support over time (Green, Gerber, & De Boef, 1999). Usually, a sample replicate is opened for data collection every day, and candidate support is estimated as 3- or 5-day MOVING AVERAGES. Rolling sample surveys use similar techniques to investigate a broader range of social science research questions, such as the dynamics of opinion change (Henry & Gordon, 2001).

—Gary T. Henry

REFERENCES

- Green, D. P., Gerber, A. S., & De Boef, S. L. (1999). Tracking opinion over time: A method for reducing sampling error. *Public Opinion Quarterly*, 63, 178–192.
- Henry, G. T., & Gordon, C. S. (2001). Tracking issue attention: Specifying the dynamics of the public agenda. *Public Opinion Quarterly*, 65(2), 157–177.
- John, K. E. (1989). The polls—A report. *Public Opinion Quarterly*, 53, 590–605.
- Kish, L. (1965). *Survey sampling*. New York: Wiley.
- Moore, D. W. (1999, June/July). Daily tracking polls: Too much “noise” or revealed insights? *Public Perspective*, pp. 27–31.
- Rhee, J. W. (1996). How polls drive campaign coverage. *Political Communication*, 13, 213–229.

SERIAL CORRELATION

Serial correlation exists when successive values of a TIME-SERIES process exhibit nonzero COVARIANCES.

The term *serial correlation* is often used synonymously with autocorrelation, which refers to the more general case of correlation with self and may be either over time, as when the president’s approval rating today is correlated with his approval rating yesterday (serial correlation), or across space, as when gang membership is correlated with physical distance of residence from an inner city (SPATIAL correlation). Spatial correlation is less pervasive than serial correlation and is seldom discussed at any length in textbook treatments of autocorrelation. Serial correlation is a common feature of time-series processes and has implications for dynamic MODEL SPECIFICATION. The presence of serial (or spatial) correlation in the errors of a REGRESSION model violates the GAUSS-MARKOV regression ASSUMPTIONS and invalidates inference from OLS estimates.

A time-series process is serially correlated if its value at time t is a LINEAR function of some number, p , of its own past values:

$$y_t = a + b_1 y_{t-1} + b_2 y_{t-2} + \cdots + b_p y_{t-p} + \mu_t.$$

If y_t is correlated only with y_{t-1} , then we say that we have first-order serial or autocorrelation, often denoted ρ_1 . If y_t is correlated with y_{t-1} and also with y_{t-2} , we have second-order autocorrelation, denoted ρ_2 , and so on. The autocorrelations themselves, the ρ , give the strength of CORRELATION. Serial correlation does not depend on the time period itself, only on the interval between the time periods—the length of the LAG—for which the correlation is being assessed. The correlation between y_t and y_t is perfect or 1 ($\rho_0 = 1$). The estimated serial correlations give us a sense of the nature and extent of the time dependence inherent in the y_t process. For first-order serial correlation, by far the most common case, the correlations at successive lags will decay at a rate determined by ρ_1 , with large correlations indicating stronger time dependence and slower decay.

If a series is covariance stationary (the mean, variance, and lagged covariances are constant over time), we can define the autocorrelations between y_t and y_{t-p} as the covariance between y_t and y_{t-p} divided by the autocovariance (variance) of y_t . Furthermore, for a stationary series, the correlations range from -1.0 to 1.0 , with larger magnitudes characteristic of stronger over-time dependence.

EXAMPLES

Typically, social processes will exhibit positive rather than negative serial correlation: Successive

observations will tend to move in the same direction—either up or down—over time. Consider some examples.

- Government or agency budgets are based on the previous budget(s) and are not created anew each year, so that changes in the process are incremental. Correlations in adjacent periods will be strong and positive.
- The aggregate partisan balance (percent of Democrats, for example) tends to move only sluggishly, presumably because features of the political system keep the balance of partisanship from varying wildly over time. This produces strong positive serial correlation in the partisan balance.
- The percent of women working in professional or managerial positions has tended to increase slowly over time as mean levels of education of women increased, producing strong positive correlations between adjacent values.

In contrast, negative serial correlation occurs when the current value of a process is related to its past in such a way that the signs on the b above oscillate. A process that adjusts over time for some sort of error in the past realization would create negative correlation, where in one time period, the process goes up, and then in the next, it goes down.

IMPLICATIONS

Serial correlation has two types of implications for MODELING. First, the analyst must be sensitive to time-series dynamics when selecting a model specification. Serial correlation in the processes we care about should be directly modeled by including lags of the dependent, and possibly independent, variables in regression models of the process. Second, analysts must also be aware that the error process in a dynamic regression may itself be serially correlated; the errors in one time period may be related to past errors in a systematic way, even after care has been taken to model dynamics directly. Most typically, the PREDICTION error is systematically too high or too low, producing positive serial correlation in the model errors. Serial correlation in the errors of a regression model violates the regression assumption that the errors are uncorrelated and invalidates inference by BIASING the estimated STANDARD ERRORS. The direction of the bias varies with the sign of the correlation, with positive correlations

deflating the standard errors and negative correlations inflating them. Given that social science processes tend to exhibit positive correlation, the variance typically will be underestimated (we inflate our estimate of the precision of our estimates), making it likely that the analyst may reject a true null hypothesis. In addition, R^2 is overestimated in this case. More generally, given serially correlated errors, t - and f -tests are no longer valid and will be misleading (Gujarati, 2002).

Given the high cost associated with drawing INFERENCES from models with serially correlated errors, good practice when estimating REGRESSION models involving time-series data includes postestimation diagnostic tests for serial correlation in the error process. The most common of these is the DURBIN-WATSON test for first-order autocorrelation.

Evidence for serial correlation in the errors may occur for several reasons. First, the evidence may be due to dynamic MISSPECIFICATION. This may occur for several reasons, such as the following:

- Excluding variables whose behavior is time dependent will produce serial correlation in the regression error; the error term will include the omitted time-dependent variable. If we omit September 11th from a model of President Bush's approval ratings, for example, we will underestimate his approval systematically after that time period. In this case, there would be strong positive autocorrelation in the errors.
- Incorrect FUNCTIONAL form can produce serial correlation. If we estimate a regression between X and Y , but Y is linearly related to the square of X , we will systematically overestimate a series of consecutive Y s and then consistently underestimate a series.
- Omitting a lagged term when the dependent variable is inertial will produce serial correlation.

Dynamic misspecification is best handled with changes in the specification of the dynamics.

Alternatively, the underlying errors themselves may be serially correlated. This may be due to problems with the data, for example. Data manipulation—INTERPOLATING MISSING values, SMOOTHING time series to deal with RANDOM fluctuations in the data (averaging consecutive observations), or aggregating over time—may lead to systematic patterns in the disturbances. In these cases, remedial measures such

as (feasible) GENERALIZED LEAST SQUARES (GLS) are required for valid inferences.

ADDITIONAL ISSUES

The discussion above focuses on serial correlation in the context of stationary time-series processes. Stationary processes represent an important class of time-series processes that are stable, having a constant mean, variance, and covariance for all time periods. Serial correlation patterns for stationary time series exhibit rapid decay—the recent values of the time series are related more strongly to its value today than are time points in the past—and the rate of decay in the strength of the correlations is predictable and quick. But serial correlation may take many forms. If a series is not stationary, serial correlations at successive lags may be equal to or even greater than 1. In the former case, the effects of the past do not decay, but cumulate over time. Serially correlated processes that do not exhibit decay in their correlations are called unit root or RANDOM WALK processes. Patterns of serial correlation that reveal stronger correlations at longer lags are called explosive processes. Both of these types of processes are examples of integrated processes. Alternatively, serial correlations may decay over time but decay so slowly that effects persist even at long lags. This class of processes is called fractionally integrated. Traditional PROBABILITY theory does not apply to non-stationary processes of any type, so that these series must be TRANSFORMED in some way to make them stationary before hypothesis tests can be conducted and inferences about social processes can be made.

—Suzanna De Boef

See also ARIMA

REFERENCE

Gujarati, D. N. (2002). *Basic econometrics* (4th ed.). New York: McGraw-Hill.

variance (or, alternatively, as the squared PEARSON'S CORRELATION COEFFICIENT between the sample values and their normalized expected standard normal-order statistics). Formally,

$$W = \frac{(\sum_{i=1}^N a_i X_i)^2}{\sum_{i=1}^N (X_i - \bar{X})^2} \quad (1)$$

where the X_i are the *ordered* values of the sample data $\mathbf{X}$ (from lowest to highest) and a_i are the normalized standard normal-order statistics $\mathbf{m}' = (m_1 \dots m_N)$ with variance-covariance matrix $\mathbf{V}$, such that $\mathbf{a}' = (a_1 \dots a_N) = \mathbf{m}'\mathbf{V}^{-1}[(\mathbf{m}'\mathbf{V}^{-1})(\mathbf{V}^{-1}\mathbf{m})]^{-1/2}$ is anti-symmetric and, by construction, $\mathbf{a}'\mathbf{a} = 1$. Under the null hypothesis of normality, W converges asymptotically to the distribution outlined in Shapiro and Wilk (1965). Low values of the test statistic support rejection of the NULL HYPOTHESIS that the data are normally distributed.

The original Shapiro-Wilk test was designed for sample sizes between 3 and 50; more recent modifications of the test have increased its usability up to N s of 2,000 (see Royston, 1993, for an overview), whereas the Shapiro-Francia W' test is valid for sample sizes up to 5,000. The test is scale and origin invariant, and is very robust to departures from non-normality; for these reasons, it is generally held to be a superior alternative to most other normality tests (e.g., the Bera-Jarque moment test and the Lilliefors modification of the KOLMOGOROV-SMIRNOV TEST).

As an example, consider the data on aggregate IQ scores from 81 countries given in Gill (2002, p. 103). The testing instrument is designed to give a mean response of 100, with a variance of 15; by contrast, the sample data have a mean of 88 and a variance of 12. The Shapiro-Wilk statistic W equals 0.944, whereas the Shapiro-Francia W' equals 0.950; both allow us to reject the null hypothesis of normality at $p < .01$.

—Christopher Zorn

SHAPIRO-WILK TEST

The Shapiro-Wilk test (typically called “ W ”) is a test for the NORMAL DISTRIBUTION of a RANDOM VARIABLE. Intuitively, this is obtained as the ratio of the square of the LINEAR COMBINATION of standard normal-order statistics to the sample

REFERENCES

- Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. London: Chapman & Hall.
- Royston, J. P. (1993). A toolkit for testing for non-normality in complete and censored samples. *The Statistician*, 42, 37–43.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.

SHOW CARD

A show card is a visual device used in standardized QUANTITATIVE face-to-face survey INTERVIEWING. It is so called because it is traditionally made of cardboard and is used to show the interviewee certain information. Typically, the information on the card consists of a list of possible answers to a question, from which the interviewee will be asked to select one or more, as appropriate. The main reason for using a show card is to overcome the difficulties that many interviewees would have retaining the response options in their head if the interviewer were simply to read them out, without providing any visual aid.

These difficulties in the absence of a show card sometimes lead to “recency effects,” whereby interviewees may tend to choose the last, or one of the later, options (i.e., the ones most recently read out) because the earlier ones are less memorable. Consequently, a show card may be used if there are many response options (say, more than three or four), or if the options are long and/or complex (e.g., phrases to describe a range of possible views on a topic). However, show cards do not overcome all problems associated with predetermined lists of response options. In some circumstances, responses may be subject to “primacy effects,” whereby interviewees tend to choose the first option that seems to apply in order to save the cognitive burden and time associated with reading the whole card in full.

Another use of a show card is to serve as a memory jogger for an OPEN-ENDED QUESTION. For example, a question asking whether the interviewee has performed any kind of voluntary activity in the past 7 days may utilize a show card listing some examples of voluntary activity in order to provide an idea of the kinds of activities in which the researcher is interested. This technique can help to prompt responses that otherwise may be omitted, but it may also lead some interviewees to focus only on listed items. In other words, interviewees may treat the list as exhaustive rather than as examples.

Show cards can also be used to display material that cannot adequately be communicated verbally. For example, an interviewee could be shown a copy of a newspaper advertisement and asked whether he or she recalls seeing it. In practice, such material acts in a way that is not very different from that of the textual memory jogger. It may prompt recognition that

would not have occurred if the interviewer had simply described the advertisement verbally, but on the other hand, it may also trigger false recognition.

The ability to use show cards (and other visual aids) is often considered to be an important advantage of face-to-face survey interviewing over telephone survey interviewing. Show cards are not free of problems, however. Unlike verbally delivered questions, they require a degree of literacy on the part of the interviewee, and this can become a source of RESPONSE BIAS. There is also the risk of primacy effects. It might be noted, however, that these issues also apply to SELF-COMPLETION surveys. The logistical aspects of the use of show cards are also important. Show cards must be produced in a way that allows easy turning from one to the next and must be (and must remain) in the correct order. (Loose-leaf cards are not a good idea.) The cards should be labeled clearly, and the interviewer should be provided with clear instructions on which card is to be used at which point in the interview (e.g., “Now look at card C . . .”).

—Peter Lynn

SIGN TEST

The sign test, in its simplest form, provides a comparison of the probabilities of two types of outcomes. The same procedure can be used to make inferences about the mean of a population of differences.

Suppose a RANDOM SAMPLE of 22 registered voters is asked whether they intend to vote for Candidate A or Candidate B in the upcoming election. Seven voters say they favor Candidate A, 12 prefer Candidate B, and three are undecided. Does this indicate a significant preference for Candidate B over Candidate A? We ignore the three undecided voters, on the basis that they are equally likely to end up voting for either candidate, and consider the remaining $n = 19$ voters with 12 “successes” for Candidate B. The test is based on this number S of successes in the sample. If the candidates are equally preferred, the PROBABILITY of a success for Candidate B from any voter would be 0.5, the same as for Candidate A. The probability of $S = 12$ or more successes for Candidate B is calculated from the BINOMIAL DISTRIBUTION as

$$P(S \geq 12) = \sum_{s=12}^{19} \binom{19}{s} (0.5)^s (0.5)^{19-s} = 0.1796.$$

This value is not particularly small, and we conclude that this sample does not provide sufficient evidence to conclude that B is significantly preferred over A. For larger sample sizes, the test is based on the NORMAL approximation to the binomial distribution (with a continuity correction), calculated from

$$Z = \frac{S - 0.5 - (n/2)}{0.5\sqrt{n}} = 0.92.$$

This normal approximation gives $P = 0.1788$, which is extremely close to the exact value.

The other type of application of the sign test is when we have a random sample of n pairs of observations $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ drawn from a bivariate distribution. We want to make an inference about the mean differences between the X s and Y s in the population. The first step is to take the differences $D_i = X_i - Y_i$ between the X and Y observation in each pair. Because the difference between the X and Y population means is the same as the mean μ_D of the $D = X - Y$ population of differences, the null hypothesis is $H_0 : \mu_D = 0$.

To carry out the sign test, we take the differences $D_1, D_2, \dots, D_n$ and note the number S of plus signs among these differences. If the alternative is $H_1 : \mu_D > 0$, the test rejects for large values of S , or equivalently, the P VALUE is in the right tail of the binomial distribution with probability of success $p = 0.5$. This application of the sign test is a special case of the BINOMIAL test of the null hypothesis that $p = p_0$ against the alternative that $p > p_0$.

As an example, suppose 10 married couples are randomly selected to determine the relationship between ratings of husbands and wives concerning the taste, convenience, and cost of frozen croissant pizza. The responses were on a scale of 0 (terrible product) to 100 (excellent product). The data are shown in Table 1. Do the husbands like the product better than their wives?

We ignore the one couple who gave the same rating. Of the remaining nine differences, eight husbands liked the product better than their wives. The probability of this occurring under the null hypothesis of equal

preference is $P(S \geq 8) = 0.0195$. With a p value this small, we reject the null hypothesis and conclude that the mean preference rating from husbands is larger than the mean rating from wives.

—Jean D. Gibbons

SIGNIFICANCE LEVEL. See ALPHA, SIGNIFICANCE LEVEL OF A TEST

SIGNIFICANCE TESTING

Significance testing provides objective rules for determining whether a researcher's HYPOTHESES are supported by the data. The need for objectivity arises from the fact that the hypotheses refer to values for POPULATIONS of interest, whereas the data come from SAMPLES of varying reliability in representing the populations. For example, the population of interest may be people of college age, and the sample may be students from a particular class who are available at the time the data are collected. This, of course, would not represent a RANDOM SAMPLE, which would be ideal, but does represent a "handy" sample often used in social science research. (Extrastatistical considerations come into play when deciding the extent to which the data from a "handy" sample generalize beyond that sample. For example, the results of a study on visual perception with a handy sample of college students should be readily generalizable.) To conclude that a particular VARIABLE influences behavior, the researcher must follow the principles of good research design in obtaining and recording the data. No amount of statistical "massaging" of the data can rescue a poorly designed study.

Even when the data come from a well-designed study, just showing that there are differences between the scores for, say, two groups is not enough. The researcher must also demonstrate that differences in behavior between the two groups are RELIABLE; that is, they are greater than differences attributable to chance.

Chance, or, more formally, chance sampling effects, refers to preexisting differences between samples of individuals that have nothing to do with how they are classified or treated in a research study. Chance sampling effects might be substantial in a study of

Table 1 Data on Preferences

Couple	1	2	3	4	5	6	7	8	9	10
Husband	76	83	60	90	100	55	62	69	75	43
Wife	71	81	62	84	65	44	61	67	75	40
Difference	+	+	−	+	+	+	+	+	0	+

human behavior where the subjects' prior history of genetic and environmental influences is beyond the researcher's control. In order to eliminate any systematic bias in the assignment of subjects to conditions based on the uncontrolled factors, researchers typically use RANDOMIZATION to assign subjects to conditions such that all subjects have the same chance of being assigned to any condition. This is a more crucial application of randomness than selecting subjects from a population that may be imprecisely defined.

Even with the use of randomization, however, a measure such as the sample MEAN can vary considerably from sample to sample even if the populations from which the samples are drawn are equivalent. This variation is due to the fact that different individuals fall into different samples, and each individual has unique characteristics. So, of course, the samples will vary one from the other. This will be especially true when sample size is low and a few discrepant scores can skew the results. Later, when we develop a specific example, we will see that sample size plays a crucial role in significance testing. To preview that illustration, the smaller the sample size, the larger the mean difference between groups has to be in order to demonstrate a statistically significant difference.

The formal process by which the researcher determines whether a difference is "statistically significant," that is, whether it represents more than chance variation, is known as HYPOTHESIS TESTING.

HYPOTHESIS TESTING

Most research questions in the social sciences boil down to two competing hypotheses. The competing hypotheses are referred to as the null hypothesis and the alternative hypothesis. The NULL HYPOTHESIS (H_0) is the hypothesis initially assumed to be true. In most cases, H_0 assumes no difference (e.g., the experimental treatment has no effect), whereas the ALTERNATIVE HYPOTHESIS (H_a) typically affirms the existence of differences among experimental conditions. The null hypothesis does not usually correspond to the actual prediction of the experiment. It is more like a "straw man" that the researcher hopes to discredit. However, because the null hypothesis is stated as an equality (e.g., the difference between the population mean scores of the experimental and control groups is zero) and the alternative hypothesis is stated as an inequality (the difference between the means is not equal to zero), only H_0 can be tested directly, and that is why we start with

that hypothesis even though it is not the actual research hypothesis. Examples of null hypotheses include the following statements: a drug has no effect; individuals will make the same decisions whether they are alone or in groups; there is no relation between income and political party affiliation; men and women receive equal pay for the same work. To affirm the existence of an effect or a RELATIONSHIP, the researcher must first reject or nullify H_0 (that is why it is called the "null" hypothesis) by providing data that are inconsistent with the assumption of no effect. Hence, the term *indirect proof* or *indirect support* has been adopted.

In testing these hypotheses, the particular statistical test used (e.g., *T-TEST*, ANALYSIS OF VARIANCE, REGRESSION, CHI-SQUARE TEST) will depend on features used to generate the data being tested. These include whether the study was designed to examine differences between experimental conditions or measure the strength of ASSOCIATIONS between sets of observations, the type of data collected (e.g., numerical data or frequency counts), and the number of factors (variables) in the EXPERIMENTAL DESIGN. The flow chart in Figure 1, reprinted from Levin (1999), illustrates the selection of tests based on these considerations (see also Howell, 2002, and Tabachnick & Fidell, 2001).

Regardless of the particular statistical test that is appropriate for a given set of data, the logic is similar. A measure is taken of the difference between conditions or degree of association that is actually observed in the sample data. This is then compared to the value expected if the null hypothesis were actually true, that is, no difference between conditions or no association between sets of observations, and only chance factors were influencing the data. *The greater the extent to which the observed differences exceed those expected by chance, the less the likelihood that the differences were due to chance and the more likely they were actually due to something else.* In a well-designed experimental study, that "something else" would be the effect of the INDEPENDENT VARIABLE manipulation on the DEPENDENT VARIABLE. In a well-designed OBSERVATIONAL study, the "something else" could be differences between categories of people such as Republicans and Democrats on behaviors such as amount of money donated to candidates for political office.

Modern computational procedures allow us to compute the PROBABILITY (the so-called *P VALUE*) that a given set of data could have been the result of chance

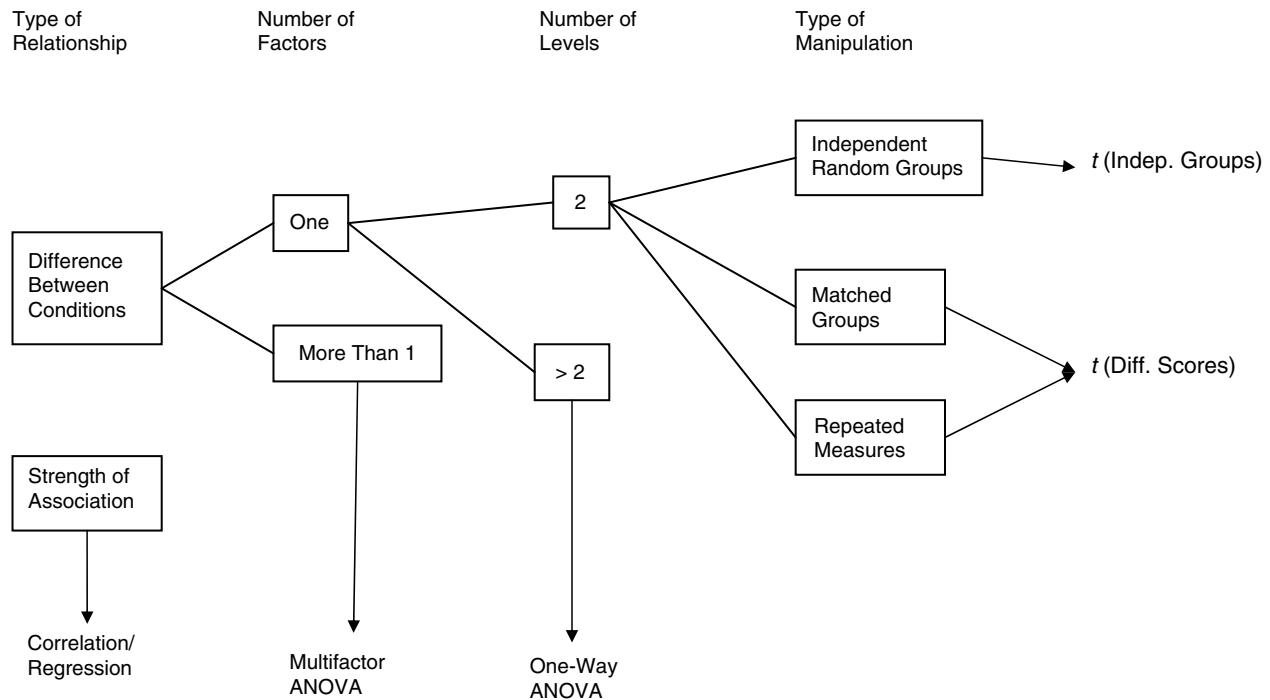


Figure 1 Selection of Tests

SOURCE: Levin, 1999.

factors alone, that is, that the null hypothesis was actually true. The lower this probability is, the less likely it is that the null hypothesis was true. So, we reject H_0 in favor of H_a when p is low and we retain H_0 when p is high. But what constitutes “low” or “high” values of p ? We had said earlier that significance testing provides objective rules for testing hypotheses. Here, some subjectivity is exercised. Clearly, if $p < .01$, we would want to reject H_0 because there is less than 1 chance in 100 that we could have gotten the data we did if the null hypothesis were true. And clearly, if $p > .25$, we would have to retain H_0 because the results could well have been due to chance. What about values in between?

These cases involve consideration of the types of errors that can be made in testing statistical hypotheses. Errors occur when H_0 is true, but you reject it in favor of H_a , or when H_0 is false, but you retain it. By convention, these two types of errors are called TYPE I ERRORS and TYPE II ERRORS, respectively. The probability of a Type I error (called ALPHA and denoted α) and the probability of a Type II error (called BETA and denoted β) are inversely related. If a strict criterion is set for rejecting H_0 , then H_0 will seldom be rejected when it is true, but H_0 is apt to be retained when it is false, resulting in a low level of alpha and a relatively high level of beta.

The opposite occurs with a lax criterion for rejecting H_0 . This criterion is called the LEVEL OF SIGNIFICANCE, which can be set ahead of time as low (e.g., $\alpha = .01$) to reduce the probability of a Type I error or high (e.g., $\alpha = .10$) to reduce the probability of a Type II error. This decision should be based on the judged relative severity of Type I and II errors. For example, in medical research, a Type I error could mean that an ineffective drug or treatment was said to have an effect, and a Type II error could mean that a helpful drug or treatment was said to have no effect. Thus, a medical researcher testing the effectiveness of a new drug might want to set $\alpha = .01$ to avoid premature marketing of an unproven drug. In practice, an intermediate level of $\alpha = .05$ is often used.

AN EXAMPLE

The t -test is most often used to compare the means of two groups or conditions. The null hypothesis is that the population mean scores are equal for the two conditions. In other words, no difference would be found if we could compare the entire population of scores for the two conditions, $H_0 : \mu_1 - \mu_2 = 0$, where μ_1 and μ_2 represent the population mean values for the two conditions. But the data, of course,

come from samples, not entire populations. In a study of aggressive behavior in children, for example, one group or condition could be a sample of heavy viewers of violence on TV, and the other could be a sample of light viewers. The alternative hypothesis is that there is a difference, $H_a : \mu_1 - \mu_2 \neq 0$. (If the direction of difference were predicted ahead of time, a ONE-TAILED version of H_a would be specified, such as $\mu_1 - \mu_2 > 0$.)

The test assumes that the distribution of scores in each condition is NORMAL and that the two distributions have equal variance. There is considerable evidence, however, that the t -test is relatively unaffected by moderate deviations from these assumptions when sample sizes are equal. The following values are computed for the two samples of scores obtained in the study: the means, $\bar{X}_1$ and $\bar{X}_2$; the STANDARD DEVIATIONS, s_1 and s_2 ; and the sample sizes, n_1 and n_2 . (Note that Greek letters are often used to denote population values, and Latin letters are used to denote sample values. This is done to clearly distinguish population from sample. Hypotheses are stated for the populations of interest; they are tested with data from samples.) The value of t is then computed using the following formula for the case of equal sample size, $n_1 - n_2 = n$:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s_1^2 + s_2^2)/n}}.$$

Note that the numerator is a measure of how much the sample means actually differed, whereas the denominator contains measures of variability that represent the influence of chance. The larger the ratio, the more likely the results represent more than chance variations.

The resulting value of t is then compared with the CRITICAL VALUE found in a table of the t statistic using the selected level of α and the appropriate DEGREES OF FREEDOM, $df = (n_1 - 1) + (n_2 - 1) = 2(n - 1)$ in our case. The critical value defines that portion of the tail or tails of the distribution of the t statistic that represents an area equal to α . This area is called the critical region. If the absolute value of t calculated from the data is larger than the value in the table, then the decision is to reject H_0 in favor of H_a . If the calculated absolute value of t is less than the value in the table, then H_0 must be retained as a viable option. The rationale in the case of retaining H_0 is that the observed difference is not sufficient in magnitude to conclude that the two conditions are different. The rationale in the case of

rejecting H_0 is that the observed difference between the means of the two samples is too great to be attributable to chance, so the researcher concludes that scores for the two groups are truly different. This is often referred to as a STATISTICALLY SIGNIFICANT difference.

To provide a numerical illustration, suppose that the mean rating by parents and teachers of each child's aggressive behavior on a scale of 1 to 7 was 5.0 for the heavy viewer group and 3.6 for the light viewer group. Further suppose that the standard deviation of the scores for each group was 2.0. To underscore the importance of sample size, let's first consider that these data came from a sample of 10 children in each group and then consider samples of size 25. In the case of samples of size 10, the value of t turns out to be 1.57, which, with degrees of freedom of 18, is not significant (does not exceed the critical value) for the traditional .05 level of significance. In the case of samples of size 25, the value of t turns out to be 2.47, which, with degrees of freedom of 48, is significant. Thus, only in the latter case would the researcher conclude that the result was statistically significant and that the null hypothesis of no effect of type of TV exposure can be rejected in favor of the alternative hypothesis that the group of children who watched much violence on TV committed more aggressive acts than the group who watched little violence on TV.

SUMMARY

SIGNIFICANCE TESTING is designed to provide a decision rule based on an estimate of the probability or likelihood that a given set of research results was due merely to chance factors varying randomly between different conditions. When the likelihood of obtaining a set of results by chance is estimated to be less than a predetermined level of significance, the most reasonable conclusion is that the results are due to more than just chance and the decision is to reject the hypothesis of no effect. If the research has been designed and conducted appropriately, the researcher may then conclude that the factor of interest—the independent variable being manipulated, for instance—had a likely effect on behavior.

—Irwin P. Levin

REFERENCES

- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

- Levin, I. P. (1999). *Relating statistics and experimental design: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–125). Thousand Oaks, CA: Sage.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Computer-assisted research design and analysis*. Boston: Allyn and Bacon.

SIMPLE ADDITIVE INDEX. See MULTI-ITEM MEASURES

SIMPLE CORRELATION (REGRESSION)

Simple CORRELATION is a measure used to determine the strength and the direction of the relationship between two variables, X and Y . A simple correlation coefficient can range from -1 to 1 . However, maximum (or minimum) values of some simple correlations cannot reach unity (i.e., 1 or -1). Given there are two sets of observations deviating from the correspondent means, $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$, a generalized simple correlation (r_{xy}) can be defined as

$$\frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}$$

or

$$\frac{\text{cov}(X, Y)}{s_X s_Y}.$$

s_X^2 and s_Y^2 are VARIANCES of the variables X and Y , and $\text{cov}(X, Y)$ is the COVARIANCE between the variables X and Y . Examples of this generalized simple correlation include Kendall's TAU coefficient, SPEARMAN CORRELATION COEFFICIENT, and PEARSON CORRELATION COEFFICIENT.

A simple correlation between variables X and Y does not imply that either variable is an INDEPENDENT VARIABLE or a DEPENDENT VARIABLE. An observed relationship between job satisfaction and job performance reveals little as to whether satisfied employees are motivated to perform well, or those who perform well are proud of themselves, which, in turn, makes them feel more satisfied with their jobs. Furthermore, a simple correlation may or may not reflect the real association between two variables because the relationship can be influenced by one or more third

variables. For instance, the observed relationship between school grade and intelligence for a group of elementary students could be influenced by students' socioeconomic status.

Assuming there is a linear relationship between the variables X and Y , we can use variable X to predict variable Y , or vice versa, with two REGRESSION equations, $Y = a_1 + b_{y.x}X + e_1$, and $X = a_2 + b_{x.y}Y + e_2$. Both $b_{y.x}$ and $b_{x.y}$ are the REGRESSION COEFFICIENTS, a_1 and a_2 are INTERCEPTS, and e_1 and e_2 are the RESIDUALS in variables X and Y that cannot be predicted by variable Y or variable X , respectively. Because $b_{y.x} = r_{xy} \frac{s_y}{s_x}$ and $b_{x.y} = r_{xy} \frac{s_x}{s_y}$, $r_{xy} = \sqrt{b_{y.x} b_{x.y}}$. Conceptually, a simple correlation in this case is a geometric mean of two regression coefficients, $b_{y.x}$ and $b_{x.y}$. Furthermore, r_{xy}^2 represents the proportion of variance accounted for by using either variable Y to predict variable X , or variable X to predict variable Y .

There are numerous simple correlations used to assess relationships with different characteristics. Some correlations are appropriate for assessing a MONOTONIC relationship (e.g., Spearman correlation coefficient), a LINEAR RELATIONSHIP (e.g., Pearson's correlation coefficient), or an ASYMMETRIC RELATIONSHIP (e.g., Somer's d). A positive or a negative monotonic relationship indicates that both variables are moving in the same direction or in opposite directions, respectively. A linear relationship, a special case of a monotonic relationship, implies that X can fairly predict variable Y (or vice versa) based on some linear function rules (e.g., $Y = a + bX$). An asymmetric relationship can occur when variable Y perfectly predicts variable X , but not vice versa.

The simple correlation can be generalized to more than two variables, such as the MULTIPLE CORRELATION COEFFICIENT or the CANONICAL CORRELATION. Whereas the former assesses the relationship between a variable (i.e., criterion) and a set of other variables (i.e., predictors), the latter measures the relationship between two sets of variables. Furthermore, simple correlation can be used for variables measured at any of the LEVELS OF MEASUREMENT. For example, we can assess the relationship between age (RATIO scale) and gender (NOMINAL scale), between Fahrenheit temperature (INTERVAL scale) and electricity consumption (ratio scale), or between letter grade (ORDINAL scale) and religion (NOMINAL scale).

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

- Bobko, P. (1995). *Correlation and regression: Principles and applications for industrial organizational psychology and management*. New York: McGraw-Hill.
- Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage.
- Hays, W. L. (1994). *Statistics* (5th ed.). New York: Harcourt Brace.

SIMPLE OBSERVATION

Simple observation techniques generate research DATA that are direct and immediate, that is, not generated or colored by interpretation, artificial EXPERIMENTAL conditions, memory, or other factors. Observational data may bear upon some perspective held by the observer, or they may be used to help develop a theory at a later time. They are often used in research areas that are not yet well developed, where careful observation can generate HYPOTHESES for future research.

Several pitfalls affect data gathering and interpretation in observational research, as described, for example, in Rosenthal (1976) and texts on social science methods (e.g., Lee, 2000). One is the question of hardware, that is, whether recording of the observed behavior is to be carried out, and whether it can be done without BIAS, without the consent or awareness of research participants, or without prohibitive ETHICAL difficulties.

Another pitfall is that observational data, if not standardized or permanently recorded for research use (and thus amenable to repeated scrutiny and VERIFICATION), may generate unreliable data when provided, for example, by multiple eyewitnesses to an event, or untrained people purportedly describing “what happened” therein. Thus, many examples exist in social science research in which differing accounts are given for the same behavior or events; these are analogous to the differing accounts and descriptions of a sporting event provided by supporters of rival teams.

A third pitfall is the fact of observer differences. That is, apart from questions of bias in interpretation, observer differences lead to inevitable VARIANCE in data. For example, we encounter variance across observers in estimations of time intervals; frequency counts; or other, presumably objective characteristics of a behavior or event. Sometimes, this problem

can be remedied by having (and later averaging) observations provided by several judges or individual observers, each observing independently according to standardized instructions, and preferably not knowing each other's identity and not being familiar with the background or purpose of the research at hand.

A fourth pitfall of simple observation relates to the possibility of the observer's also being himself or herself a covert participant in the events being observed, sometimes with this participation remaining unknown to those whose behavior is being observed. Rosenhan (1973), for example, arranged for a team of PARTICIPANT OBSERVERS (“pseudopatients”) to be admitted separately and independently to a series of 12 different U.S. psychiatric hospitals, their admissions being unknown to the real patients and hospital staff. Such procedures have the benefit of gathering objective and factual observations (and inside information) about behavior in particular settings, in addition to other information more colored by interpretive bias or theoretical preference. Clearly, the procedures also bear ethical complications—to some degree, these represent a necessary cost that must be borne in exchange for being able to carry out ecologically valid observational research unaffected by participants' knowledge and awareness.

Observational procedures, especially when enhanced by permanent recording, thus have their greatest role in generating pure, immediate data in new or uncharted research areas. The RELIABILITY and interpretation of such data (and ideas generated therein for future research) can then be assessed, verified, and expanded upon in additional research.

—Stewart Page

REFERENCES

- Lee, R. (2000). *Unobtrusive methods in social research*. Buckingham, UK: Open University Press.
- Rosenhan, D. (1973). On being sane in insane places. *Science*, 179, 250–258.
- Rosenthal, R. (1976). *Experimenter effects in behavioral research* (rev. ed.). New York: Irvington.

SIMPLE RANDOM SAMPLING

Simple random sampling (SRS) is a special case of RANDOM SAMPLING. It involves selecting units by a

random chance mechanism, such that every unit has an equal and independent chance of being selected. Thus, the two features that distinguish SRS from other forms of random sampling are the equality and the independence of the selection probabilities. In consequence, each possible combination of n distinct units out of a population of N units will have the same chance of constituting the selected sample.

Selecting balls from a well-mixed bag is often considered a good approximation to SRS. Mechanisms that are constructed for the selection of lottery numbers are designed to be practically equivalent to SRS. Simple random sampling is rarely used for social surveys because there are typically good reasons to incorporate into the design features such as stratification (see STRATIFIED SAMPLE), clustering (see MULTISTAGE SAMPLING), and disproportionate sampling fractions (see SAMPLING FRACTION). However, it does provide a useful benchmark against which alternative sample designs can be compared, particularly in terms of BIAS and in terms of the VARIANCE of ESTIMATORS (see DESIGN EFFECTS).

Simple random sampling typically features heavily in sampling textbooks and courses because it provides a simple case for which the theory and algebra can be developed. This can then be extended to cover other, more realistic, sample designs.

—Peter Lynn

REFERENCES

- Barnett, V. (1974). *Elements of sampling theory*. Sevenoaks, UK: Hodder & Stoughton.
- Scheaffer, R. L., Mendenhall, W., & Ott, L. (1990). *Elementary survey sampling*. Boston: PWS-Kent.

SIMULATION

Most social science research either develops or uses some kind of theory or model; for instance, a theory of cognition or a model of the class system. Generally, such theories are stated in textual form, although sometimes, the theory is represented as an equation (e.g., in REGRESSION ANALYSIS). As personal computers became more common in the 1980s, researchers began to explore the possibilities of expressing theories as computer programs. The advantage is that social processes then can be simulated in the computer, and in some circumstances, it is even possible to carry out

experiments on artificial social systems that would be quite impossible or unethical to perform on human populations.

The earliest attempts to simulate social processes used paper and pencil (or card tiles on a checker board), but with the advent of powerful computers, more powerful computer languages, and the greater availability of data, there has been an increased interest in simulation as a method for developing and testing social theories.

The logic underlying the methodology of simulation is that a model is constructed (e.g., in the form of a computer program or a regression equation) through a process of abstraction from what are theorized to be the actually existing social processes. The model is then used to generate expected values, which are compared with empirical data. The main difference between statistical modeling and simulation is that the simulation model can itself be “run” to produce output, whereas a statistical model requires a statistical analysis program to generate expected values. As this description implies, simulation tends to imply a REALIST philosophy of science.

THE ADVANTAGES OF SIMULATION

To create a useful simulation model, its theoretical presuppositions need to have been thought through with great clarity. Every relationship to be modeled has to be specified exactly, and every parameter has to be given a value—otherwise, it will be impossible to run the simulation. This discipline means that it is impossible to be vague about what is being assumed. It also means that the model is potentially open to inspection by other researchers in all its detail. These benefits of clarity and precision also have disadvantages, however. Simulations of complex social processes involve the estimation of many parameters, and adequate data for making the estimates can be difficult to obtain.

Another benefit of simulation is that it can, in some circumstances, give insights into the “emergence” of macro-level phenomena from micro-level action. For example, a simulation of interacting individuals may reveal clear patterns of influence when examined on a societal scale. A simulation by Nowak and Latané (1994), for example, shows how simple rules about the way in which one individual influences another’s attitudes can yield results about attitude change at the level of a society, and a simulation by Axelrod (1995) demonstrates how patterns of political domination

can arise from a few rules followed by simulated nation-states. Schelling (1971) used a simulation to show that high degrees of residential segregation could occur even when individuals were prepared to have a majority of people of different ethnicity living in their neighborhood.

In this simulation, people's homes are represented by squares on a large grid. Each grid square is occupied by one simulated household (either a "black" or a "white" household) or is unoccupied. When the simulation is run, each simulated household in turn looks at its eight neighboring grid squares to see how many neighbors are of its own color and how many are of the other color. If the number of neighbors of the same color is not sufficiently high (e.g., if there are fewer than three neighbors of its own color), the household "moves" to a randomly chosen unoccupied square elsewhere on the grid. Then, the next household considers its neighbors and so on, until every household comes to rest at a spot where it is content with the balance of colors of its neighbors.

Schelling noted that when the simulation reaches a stopping point in which households no longer wish to move, there is always a pattern of clusters of adjacent households of the same color. He proposed that this simulation mimicked the behavior of whites fleeing from predominantly black neighborhoods, and he observed from his experiments with the simulation that even when whites were content to live in locations where there were a majority of black neighbors, the clustering still developed: Residential segregation could occur even when households were prepared to live among those of the other color.

AGENT-BASED SIMULATION

The field of social simulation has come to be dominated by an approach called *agent-based simulation* (also called *multiagent simulation*). Although other types of simulation, such as those based on system dynamics models (using sets of difference equations) and *MICROSIMULATION* (based on the simulated aging of a survey sample to learn about its characteristics in the future) are still undertaken, most simulation research now uses agents.

Agents are computer programs (or parts of programs) that are designed to act relatively autonomously within a simulated environment. An agent can represent an individual or an organization according to what is being modeled. Agents are generally

programmed to be able to "perceive" and "react" to their situation, pursue the goals they are given, and interact with other agents, for example, by sending them messages. Agents are generally created using an object-oriented programming language and are constructed using collections of condition-action rules.

Agent-based models have been used to investigate the bases of leadership, the functions of norms, the implications of environmental change on organizations, the effects of land-use planning constraints on populations, the evolution of language, and many other topics. Examples of current research can be found in the *Journal of Artificial Societies and Social Simulation*, available at <http://www.soc.surrey.ac.uk/JASSS/>.

Although most agent-based simulations have been created to model real social phenomena, it is also possible to model situations that could not exist in our world, in order to understand whether there are universal constraints on the possibility of social life (e.g., Can societies function if their members are entirely self-interested and rational?). These are at one end of a spectrum of simulations ranging from those of entirely imaginary societies to those that aim to reproduce specific settings in great detail.

A variant on agent-based modeling is to include people in place of some or all of the computational agents. This transforms the model into a type of multi-player computer game, which can be valuable for allowing the players to learn more about the dynamics of some social setting (e.g., business students can be given a game of this type in order to learn about the effects of business strategies). Such games are known as participatory simulations.

THE POTENTIAL OF SIMULATION

Although computer simulation can be regarded as simply another method for representing models of social processes, it encourages a theoretical perspective that emphasizes emergence, the search for simple regularities that give rise to complex phenomena, and an evolutionary view of the development of societies. This perspective has connections with complexity theory, an attempt to locate general principles applying to all systems and that show autonomous behavior—including not only human societies, but also biological and physical phenomena.

—Nigel Gilbert

REFERENCES

- Axelrod, R. (1995). A model of the emergence of new political actors. In N. Gilbert & R. Conte (Eds.), *Artificial societies*. London: UCL.
- Carley, K., & Prietula, M. (Eds.). (1994). *Computational organization theory*. Hillsdale, NJ: Lawrence Erlbaum.
- Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up*. Cambridge: MIT Press.
- Gilbert, N., & Troitzsch, K. G. (1999). *Simulation for the social scientist*. Milton Keynes, UK: Open University Press.
- Nowak, A., & Latané, B. (1994). Simulating the emergence of social order from individual behaviour. In N. Gilbert & J. Doran (Eds.), *Simulating societies: The computer simulation of social phenomena* (pp. 63–84). London: UCL.
- Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1, 143–186.

SIMULTANEOUS EQUATIONS

Many THEORIES specify relationships between variables that are more complicated than can be modeled in a single equation. Therefore, systems of simultaneous equations specify multiple equations, where each equation relates to the others. Possible relations include causal chains between DEPENDENT VARIABLES, RECIPROCAL RELATIONSHIPS between dependent variables, or CORRELATIONS between ERRORS in the equations. Simultaneous equation models can be considered a subset of STRUCTURAL EQUATION models that consider only observed variables.

Two major classes of simultaneous equations are RECURSIVE systems, where the direction of influence flows in a single direction, and NONRECURSIVE systems, where reciprocal causation, feedback loops, or correlations between errors exist. Simultaneous equation models can be specified in any of three ways: as a series of equations, as a PATH DIAGRAM, or with matrices of coefficients.

Equations in simultaneous equation systems are specified as in the following recursive simultaneous equation system:

$$y_1 = \gamma_{11}x_1 + u_1$$

$$y_2 = \beta_{21}y_1 + \gamma_{21}x_1 + \gamma_{22}x_2 + u_2$$

$$y_3 = \beta_{31}y_1 + \beta_{32}y_2 + \gamma_{32}x_2 + u_3$$

where $\text{COV}(u_1, u_2) = \text{COV}(u_1, u_3) = \text{COV}(u_2, u_3) = 0$.

In the system of equations, ENDOGENOUS VARIABLES are denoted with y s, and EXOGENOUS VARIABLES

are denoted with x s. β s indicate the effect of an endogenous variable on another endogenous variable, and γ s indicate the effect of an exogenous variable on an endogenous variable. The u s indicate errors in the equations.

After a model is specified, we must make sure it avoids the IDENTIFICATION PROBLEM. That is, we must establish whether the parameters of the model are estimable. A variety of rules and conditions to establish identification are available, including the null-beta rule, whereby models that do not have beta matrices, such as seemingly unrelated regression models, are identified; the recursive rule, which identifies all recursive models; the order condition; and the rank condition (see Bollen, 1989, pp. 98–103). Rigdon (1995) provides a straightforward way to identify many types of models that do not fall under the above rules.

Estimation of recursive systems of equations is straightforward and can be performed using ORDINARY LEAST SQUARES (OLS) regression on each equation. Estimation of nonrecursive systems of equations requires that estimation take account of the information from other equations in the system. Therefore, if OLS regression is applied to a nonrecursive simultaneous equation system, the estimates are likely to be BIASED and inconsistent. A variety of alternative estimation techniques are possible for nonrecursive simultaneous equation models, including two-stage least squares, three-stage least squares, and MAXIMUM LIKELIHOOD ESTIMATION. It is possible to view all of these estimators as INSTRUMENTAL VARIABLE estimators.

—Pamela Paxton

REFERENCES

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Greene, W. H. (1997). *Econometric analysis* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Heise, D. R. (1975). *Causal analysis*. New York: Wiley.
- Rigdon, E. E. (1995). A necessary and sufficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research*, 30, 359–383.

SKewed

A variable that has a skewed distribution is one that has most of its values clustered at the extreme end of a DISTRIBUTION of values for the variable. The measures

of CENTRAL TENDENCY for the variable are not equal. The consequences of skewness can be significant, especially for ORDINARY LEAST SQUARES (OLS) regression analysis.

To illustrate, envision a measure of democracy on a 0 (not democratic) to 100 (completely democratic) scale, constructed by making judgments about how well countries score on categories that fit a conceptualization of democracy (i.e., respect for civil rights and liberties, conducting free and fair elections). If the countries in the Northern Hemisphere are sampled, then the result will be that the OECD countries will cluster at the high democracy scores from 90–100. This is a negative skew on democracy, which means that the extreme values of the distribution are to the left of 90–100 along a continuum. If the countries in the Southern Hemisphere are sampled, where there are many more repressive dictatorships, then the result is that many countries will score at the 0–10 levels. This is a positive skew, meaning that the extreme values of the distribution are to the right of 0–10 along a continuum.

Extremely skewed distributions force social scientists to be cautious in interpreting statistical analyses, especially OLS regression. One of the OLS regression assumptions is that the INDEPENDENT VARIABLES have a linear relationship to the DEPENDENT VARIABLE. Another assumption is that the error term is distributed normally, which defaults in large samples ($N > 30$) to the dependent variable being distributed normally. Dependent variables with skewed distributions are likely to lead to a violation of the normality assumption. Independent variables with skewed distributions are unlikely to lead to a BLUE (BEST LINEAR UNBIASED ESTIMATION), at least if the sample is small.

The statistic for determining the skewness of any variable is as follows:

$$\text{Skewness of } y = \left\{ \sum [(y_i - \text{mean of } y)/s_y]^3 \right\} / N,$$

where $\sum$ = sum, s = standard deviation, and N = number of observations of the variable. The more NORMAL the distribution, the closer to zero the skewness statistic will be.

Remedies exist for skewness. Standard correction techniques involve arithmetic transformations of the variable. The most common transformation is a LOGARITHMIC transformation. It is effective in increasing the dispersion of a variable, breaking up clusters around a small number of values. Other transformations that

the analyst can try include a parabolic transformation, which involves squaring the variable, or a hyperbolic transformation, which is the operation $1/y$. In some instances, however, there is no practical remedy, and the researcher will simply have to live with a non-normally distributed variable.

—Ross E. Burkhart

REFERENCE

- Weisberg, H. F. (1992). *Central tendency and variability* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–083). Newbury Park, CA: Sage.

SKEWNESS. See KURTOSIS

SLOPE

The equation for a straight line is $Y = A + BX$, where Y is the dependent variable, A is the y-intercept, B is the slope, and X is the independent variable. The slope is calculated by taking any two points on the line and dividing the rise by the run, where the run is the increase in X between the two points and the rise is the increase (or decrease) in Y between the two points (see Figure 1). A positive relationship between the two variables is indicated by a positive slope; a negative relationship is indicated by a negative slope.

When REGRESSION ANALYSIS is used in the social sciences, the data never fit perfectly along a straight line, meaning that we have error in prediction. The usual simple regression equation is $Y = A + BX + E$, where E represents the error associated with predicting Y given X . The slope is now interpreted as the average rate by which the dependent variable changes with a 1-unit change in the independent variable. In simple regression (i.e., where there is only one independent variable), the slope can tell us the marginal relationship between income and education. By way of an example, assume that we regress income (measured in dollars) on education (measured in years of schooling), obtaining a slope of 5000. We would argue, then, that an increase of one year of education is associated with, on average, an increase in income of \$5,000.

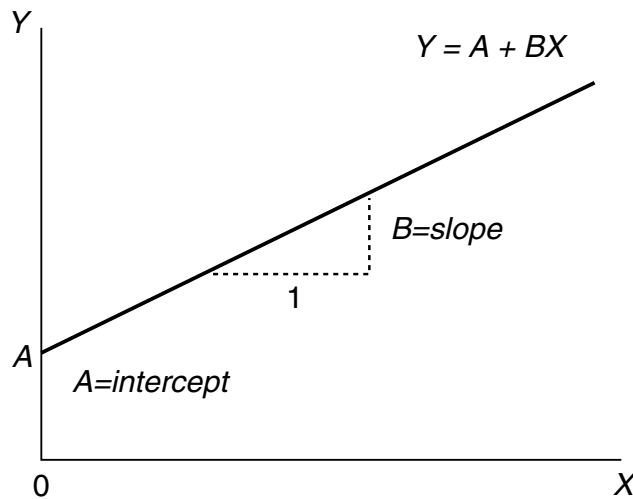


Figure 1 Diagram Depicting the Slope of a Straight Line

In MULTIPLE REGRESSION ANALYSIS (i.e., when there are several independent variables), the individual slope estimates for each independent variable pertain to partial relationships. In other words, each slope refers to the effect of a single independent variable on the dependent variable controlling for all other independent variables in the model. For example, assume that we regress income (again measured in dollars) on years of education and gender, obtaining a slope for education of 4000. The interpretation of the slope would then be the following: Holding gender constant, with every additional year of education, income increases, on average, by \$4,000.

Reporting the slope coefficients is usually sufficient when the functional form between two quantitative variables is linear because it is easily interpreted. If the functional form of the relationship between two variables is not linear (i.e., the slope of the regression line is not constant through the range of the independent variable), however, a single slope coefficient cannot summarize the relationship. In such cases, several slope estimates are needed, and graphs are usually required in order to fully comprehend the relationship. For example, in polynomial regression (see POLYNOMIAL EQUATION), the independent variable has one slope coefficient more than the number of changes in the direction of the line. In nonparametric regression (e.g., LOCAL REGRESSION), there are no slope coefficients, and the relationship can be displayed only in a graph.

The concept of slope can also be generalized to extensions of the linear model, such as GENERALIZED LINEAR MODELS and MULTILEVEL ANALYSIS. In nonlinear models such as LOGISTIC REGRESSION, slopes are usually best interpreted by calculating fitted probabilities.

—Robert Andersen

REFERENCES

- Allison, P. D. (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Pine Forge.
 Lewis-Beck, M. S. (1989). *Applied regression: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-022). Newbury Park, CA: Sage.

SMOOTHING

Smoothing can be broadly considered a tool for summarizing or drawing out the important features of data. Such a useful tool has many applications in the analysis of data. Smoothing may be employed simply to create a HISTOGRAM that reveals more detail about the distribution of a CONTINUOUS variable in contrast to a histogram with bins. This same smoothing process is also used to create the FUNCTIONAL form of the DISTRIBUTION in NONPARAMETRIC estimation. When considering BIVARIATE relationships, smoothing is a valuable mechanism to discern the functional form of the relationship in the case of NONLINEARITY (see LOCAL REGRESSION). In TIME-SERIES analysis of a single variable (UNIVARIATE), smoothing may be performed during exploratory analysis to summarize the long-run changes in the variable. Smoothing is particularly useful for the FORECASTING of future values of a time series such as inflation or growth rates. Given an erratic distribution of data points, whether due to the WHITE NOISE component of a variable measured over time, abrupt changes in the density distribution of a variable, or a nonlinear bivariate relationship, smoothing techniques convey useful information by drawing out the systematic variation or the pattern of the data.

HISTOGRAM SMOOTHING

Smoothing of a histogram, rather than the use of bins, better reveals the particularities of a continuous variable's distribution. Bins suffer from the

disadvantages of concealing the continuity of the variable's distribution and possibly misrepresenting the data as a result of the arbitrariness of bin width and bin origin. As a routine analytical practice, smoothing permits better inspection of the distribution and evaluation of the NORMALITY assumption that may be made in the use of particular ESTIMATORS.

Given a non-normal distribution of continuous data, histogram smoothing is also used in nonparametric estimation to create the functional form of the distribution. Nonparametric estimation is often employed when there is reason to believe that the assumptions of parametric estimators are not satisfied, such as a non-normal distribution. Absent theoretical SPECIFICATION of the underlying distribution of the variable, the sample data provide the functional form of the distribution.

For both the routine analysis of a variable's distribution and the detection of the functional form of the distribution in nonparametric estimation, a smoothing technique obtains the best representation of the shape of the distribution. Essentially, for a variable X , smoothing utilizes the points close to any particular value of the variable x_i to create the density of the histogram. The degree of smoothing is positively related to the magnitude of the smoothing parameter (also known as the "bandwidth"). The smoothing parameter is the proportion of observations to which the smoothing operation is applied. A relatively small value of the smoothing parameter results in less smoothing because the observations relatively closer to x_i are weighted more heavily relative to observations further away. Conversely, a relatively large value of the smoothing parameter translates into greater smoothing as the weight given to close observations relative to further observations moves toward unity. Care should be taken to avoid "undersmoothing" or "oversmoothing" because the result is the introduction of BIAS. Undersmoothing, also known as "overfitting," occurs when excessive noise remains in the data so that the histogram does not reveal the meaningful features of the distribution. Oversmoothing, or "underfitting," refers to a histogram that conceals the important features of the data. A useful exercise is to experiment with various smoothing parameters, ensuring that the selected distribution is not highly sensitive to small changes in the magnitude of the smoothing parameter.

BIVARIATE OR SCATTERPLOT SMOOTHING

Smoothing is also useful in the inspection of the functional form of a bivariate relationship. Typically, a SCATTERPLOT is generated to detect the functional form of the bivariate relationship (LINEAR, CURVILINEAR, etc.). Unfortunately, the functional form is usually unclear. Smoothing the series of data points summarizes "how the central tendency of the Y variable's distribution changes at different locations within X 's distribution" (Jacoby, 1997, p. 63). As in the detection of the functional form of a single variable's distribution in nonparametric estimation, the purpose of smoothing a bivariate distribution is to allow the data to specify the functional form of the relationship. For a description of a common bivariate smoothing method, see the entry for LOCAL REGRESSION. Again, care should be taken to avoid oversmoothing or undersmoothing. Jacoby (1997) offers some useful strategies to avoid these errors. For example, a cluster of points through which the density line does not pass or remaining structure in the smoothed residuals indicates oversmoothing (or underfitting). Figure 1 provides examples of undersmoothing, oversmoothing, and a properly smoothed bivariate distribution (where, for a hypothetical class, X = grade in percent and Y = student evaluation of professor).

TIME-SERIES SMOOTHING

When smoothing the time series of a variable, the data are essentially treated as that part consisting of the true underlying pattern plus RESIDUAL noise, or the "smooth" plus the "rough." As with smoothing in general, the objective is to draw out the true underlying pattern of the time series without oversmoothing (when the variation of the underlying pattern of the time-series is lost to excessive smoothing) or undersmoothing (a time series that exhibits excessive noise so that the underlying pattern cannot be distinguished from the noise). In both cases, the underlying pattern is not easily discernible. Oversmoothing or undersmoothing is manifested in time-series forecasting by a large prediction error.

There are various smoothing techniques for univariate time-series data (Jacoby, 1997; Mills, 1990). In part, the desirable technique depends on the purpose of analysis. *Exponential smoothing*, for example, may be desirable for forecasting because this

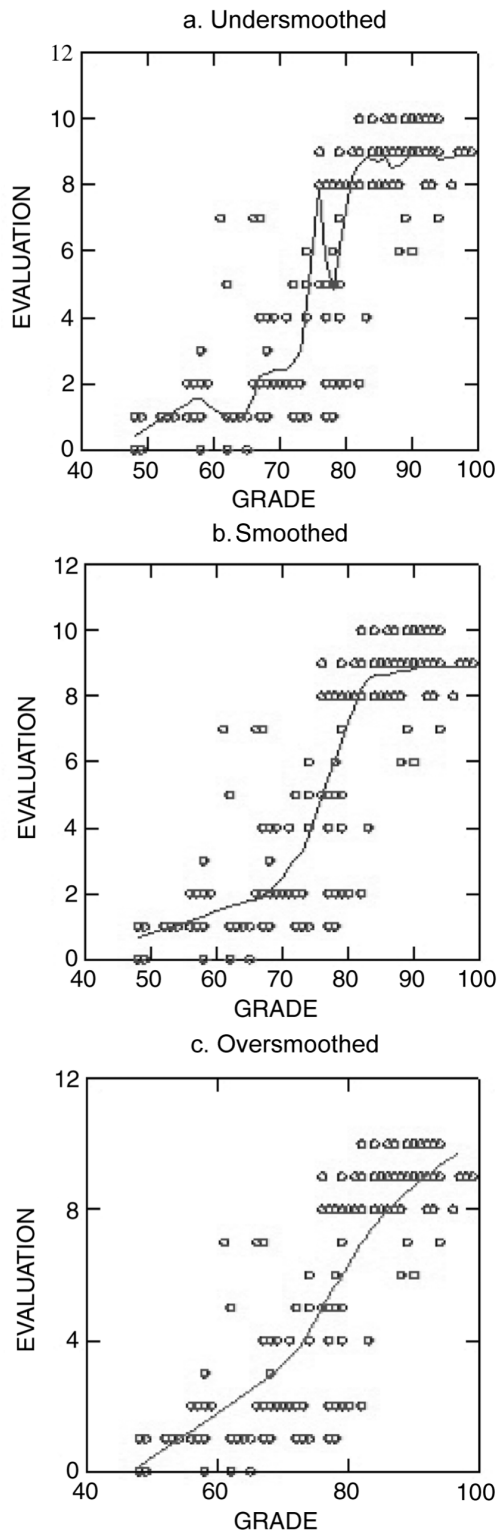


Figure 1 Smoothing Examples

technique weights recent observations more heavily than distant observations. For exploratory purposes, a simple method such as *running medians* or *running weighted averages* may suffice. Essentially, these techniques smooth the time series by replacing each data point with the median or weighted average of close data points. The minimum order, or number of close data points used to adjust each observation, is 3. For example, the formula to determine the smoothed value of the data point x_t using running weighted averages with an order of 3 is the following:

$$s_t = (w_i)x_{t-1} + (w_i)x_t + (w_i)x_{t+1},$$

where s_t is the smoothed value and w_i is the weight assigned to each observation (symmetric around x_t). The application of this formula is then repeated for every x value. The appropriate order depends, in part, on the desired smoothing magnitude.

A greater degree of smoothing also may be achieved by using repeated iterations of the same smoothing technique or sequentially applying different smoothing techniques. Of course, the larger the order and/or the greater the number of smoothing treatments, the greater the susceptibility to oversmoothing.

—Jacque L. Amoureux

REFERENCES

- Jacoby, W. J. (1997). *Statistical graphics for univariate and bivariate data*. Thousand Oaks, CA: Sage.
- Kennedy, P. (1998). *A guide to econometrics* (4th ed.). Cambridge: MIT Press.
- Mills, T. C. (1990). *Time series techniques for economists*. Cambridge, UK: Cambridge University Press.

SNOWBALL SAMPLING

Snowball sampling may be defined as a technique for gathering research subjects through the identification of an initial subject who is used to provide the names of other actors. These actors may themselves open possibilities for an expanding web of contact and inquiry. The strategy has been utilized primarily as a response to overcome the problems associated with understanding and SAMPLING concealed populations such as the deviant and the socially isolated (Faugier &

Sargeant, 1997). Snowball sampling can be placed within a wider set of methodologies that takes advantage of the social networks of identified respondents, which can be used to provide a researcher with an escalating set of potential contacts.

Snowball sampling is something of a misnomer for a technique that is conventionally associated more often with QUALITATIVE RESEARCH and acts as an expedient strategy to access hidden populations. The method has suffered an image problem in the social sciences given that it contradicts many of the assumptions underpinning conventional notions of random selection and representativeness. However, the technique offers advantages for accessing populations such as the deprived, the socially stigmatized, and the elite. In his study of political power and influence in South London, Saunders used a reputational method and asked his contacts to refer him to others who were viewed as holding power in the area (Saunders, 1979). This yielded important but also potentially self-contained systems of social contact but acted as a valuable insight into the perceptual basis of political power.

Snowball sampling has also advanced as a technique, and the literature contains evidence of a trend toward more sophisticated methods of sampling frame and error estimation and growing acceptance as a tool in the researcher's kit. Shaw, Bloor, Cormack, and Williamson (1996) illustrate these issues in their use of snowball sampling to estimate the prevalence of hard-to-find groups such as the homeless. Here, a technique viewed as a limited qualitative tool is highlighted as a systematic investigative procedure with new applications in the social sciences. However, although selection bias may be addressed through the generation of larger sample sizes and the REPLICATION of results, there also remain problems of gatekeeper BIAS, whereby contacts shield associates from the researcher. However, other groups may themselves consist of highly atomized and isolated individuals whose social network is relatively impaired, further problematizing contact with others.

Snowball sampling can be applied for two primary purposes: first, as an informal method to reach a target population where the aim of a study may be exploratory or novel in its use of the technique, and second, as a more formal methodology for making inferences with regard to a population of individuals that has been difficult to enumerate through the use of descending methodologies such as household surveys.

Early examples of the technique can be seen in classics such as Whyte's *Street Corner Society*, which used initial contacts to generate contexts and encounters that could be used to study the gang dynamic (although there has been a general move away from PARTICIPANT OBSERVATION of this kind toward the use of snowball sampling techniques for interview-based research more recently).

Although some have viewed snowball strategies as trivial or obscure, their main value lies as a method for dealing with the difficult problem of obtaining respondents where they are few in number or where higher levels of trust are required to initiate contact. Under these circumstances, such techniques of "chain referral" may imbue the researcher with characteristics associated with being an insider or group member, which can aid entry to settings in which conventional approaches have great difficulty.

—Rowland Atkinson and John Flint

REFERENCES

- Faugier, J., & Sargeant, M. (1997). Sampling hard to reach populations. *Journal of Advanced Nursing*, 26, 790–797.
- Saunders, P. (1979). *Urban politics: A sociological interpretation*. London: Hutchinson.
- Shaw, I., Bloor, M., Cormack, R., & Williamson, H. (1996). Estimating the prevalence of hard-to-reach populations: The illustration of mark-recapture methods in the study of homelessness. *Social Policy and Administration*, 30(1), 69–85.
- Whyte, W. F. (1955). *Street corner society: The social structure of an Italian slum* (2nd ed.). Chicago: University of Chicago Press.

SOCIAL DESIRABILITY BIAS

With regard to people's reports about themselves, social desirability is the tendency to respond to questions in a socially acceptable direction. This RESPONSE BIAS occurs mainly for items or questions that deal with personally or socially sensitive content. Social desirability is considered a personality variable along which individuals vary (Crowne & Marlowe, 1964), and it is only certain individuals who exhibit this bias. Crowne and Marlowe consider social desirability to be an indication of an individual's need for approval that results in a response bias rather than a response bias per se.

Social desirability as a personality variable is assessed most often with the Crowne and Marlowe (1964) social desirability scale. The scale consists of true/false questions about the self, each reflecting something that is either desirable or undesirable (e.g., "I always check out the qualifications of candidates before voting"). Considering the extreme wording "I always" or "I have never," it is almost impossible for more than a few of the items to literally be true of a person. Scoring high reflects the person's attempt, whether intentional or not, to put himself or herself in a favorable light and appear socially acceptable. Other scales also exist to assess social desirability. For example, the Paulhus (1984) scale assesses two dimensions of social desirability: other deception and self-deception.

Despite widespread concerns, there is little evidence to suggest that social desirability is a universal problem in research that relies on self-reports. For example, Moorman and Podsakoff (1992) studied the prevalence and effects of social desirability in organizational research. They concluded that relations tended to be quite small, and that social desirability produced little distortion of results among other variables. However, there are domains in which social desirability has been a potential problem, such as the area of psychological adjustment. This area is concerned with internal psychological states (e.g., anxiety and self-esteem) that are seen as socially desirable and can best be assessed through self-reports.

There are two approaches to dealing with social desirability. One is to design instruments and methods to reduce its impact on assessment. This might be accomplished through the careful choice of scale items that are not related to social desirability, which is possible for some but not all constructs of interest. The second is to assess social desirability independent of the other variables and then control for it. This could be done by eliminating from analysis individuals who score high on social desirability, or by STATISTICALLY CONTROLLING (partialing) the measure of social desirability from the other measures. This strategy can cause its own distortions, because one cannot be certain that relations of social desirability with other variables are due solely to BIAS or to real relations between the personality variable of approval need with the other variables of interest.

—Paul E. Spector

REFERENCES

- Crowne, D., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: Wiley.
- Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology*, 65, 131–149.
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598–609.

SOCIAL RELATIONS MODEL

The social relations model (Kenny & La Voie, 1984) is for dyadic or relational data from two-person interactions or group members' ratings of one another. Unlike SOCIOMETRY studies, the LEVEL OF MEASUREMENT is interval (e.g., 7-point scales) and not categorical (e.g., yes/no). Generally, the data are collected from people, but the units can be something else (e.g., organizations). The model can be viewed as a complex MULTILEVEL model.

The most common social relations design is the round-robin research design. In this design, each person interacts with or rates every other person in the group, and data are collected from both members of each dyad. Usually, there are multiple round robins. Other designs are possible. The minimum condition is that each person rate or interact with multiple others, and each person is interacted with or rated by multiple others.

There are three major types of effects in the social relations model. The actor effect represents a person's average level of a given behavior in the presence of a variety of partners. For example, Tom's actor effect on the variable of trust measures the extent to which he tends to trust others in general. The partner effect represents the average level of a response that a person elicits from a variety of partners. Tom's partner effect measures the extent to which other people tend to trust him. The relationship effect represents a person's behavior toward another individual, in particular, above and beyond the actor and partner effects. For example, Tom's relationship effect toward Mary on the variable of trust measures the extent to which he trusts her, controlling for his general tendency toward trusting others and her general tendency to be

trusted by others. Relationship effects are directional or asymmetric, such that Tom may trust Mary more or less than she trusts him. To differentiate relationship from error variance, multiple indicators of the construct, either across time or with different measures, are necessary.

The focus in social relations modeling is not on estimating the effects for specific people and relationships but in estimating the variance due to effects. So, for a study of how intelligent people see each other, the interest is in whether there is actor, partner, and relationship variance. Actor variance would assess if people saw others as similar in terms of intelligence, partner variance would assess whether people agree with each other in their ratings of intelligence, and relationship variance would assess the degree to which perceptions of intelligence are unique to the dyad.

The social relations model also permits a detailed study of reciprocity. For instance, it can be studied whether actors who tend to like others are liked by others (generalized reciprocity) and whether there is reciprocity of relationship effects (dyadic reciprocity).

There has been extensive research using the model. The most concentrated effort has been in the area of interpersonal perception (Kenny, 1994). Also investigated are the topics of reciprocity of disclosure, small-group process, and individual differences in judgmental accuracy.

—David A. Kenny

REFERENCES

- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford.
- Kenny, D. A., & La Voie, L. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 18, pp. 142–182). Orlando, FL: Academic Press.

SOCIOGRAM

A sociogram is the representation, through a graph, of relationships between individuals (or whatever the unit of analysis might be). Sociograms are helpful for mapping structures of relationships. They are usually the outcomes of research conducted under the umbrella of NETWORK ANALYSIS or SOCIOMETRY.

—Alan Bryman

SOCIOMETRY

This term is occasionally employed to refer to forms of measurement in the social sciences that are concerned with measurement through SCALES. Its more common use is in relation to an approach to research that emphasizes the mapping of relationships between units of analysis (individuals, groups, organizations, etc.). Increasingly, sociometry is called NETWORK ANALYSIS, and most research that was formally referred to as sociometric is subsumed by this field.

—Alan Bryman

SOFT SYSTEMS ANALYSIS

Soft systems analysis (SSA) is a method for investigating problems and planning changes in complex systems. It can also be used to design new systems. It was developed initially by Peter Checkland from Lancaster University (Checkland, 1981). The method can best be seen as a practical working tool, although it has been widely used in applied research. A core issue surrounds the meaning of “soft” as opposed to “hard” systems. The central idea in soft systems thinking is that people see and interpret the world differently. Discrepancies in the views held by individuals are not sources of invalidity or “noise” in the data; rather, differentiation reflects the nature of reality. Pluralism is the norm. In complex systems, individuals or groups are likely to construct quite different views of how the system works, what may be wrong with it, and how it should be improved.

Although recognizing that SSA can be used in a variety of ways, it is easiest to describe in its most straightforward format (Naughton, 1984). The method is organized in a series of relatively formal and well-structured stages through which its users work. In practice, considerable iteration can take place between the stages. The analyst initially gathers data about the problem situation (Stage 1), and this is represented in pictorial form (a “rich picture”) (Stage 2). The users of the method (i.e., the analyst and the system participants) then explicitly try looking at the system in a number of different ways, searching for views that add some new light (a “relevant system”) (Stage 3). They

select a new perspective on the problem situation and develop a conceptual model of what the system would logically have to do to meet the requirements of this new view (Stage 4). This model is then compared with the existing problem situation to see if there are any lessons for change (Stage 5). These are then discussed by the participants, who decide which should be implemented (Stage 6). Changes that are agreed to are then implemented (Stage 7). If the new view (from Stage 3) does not appear to offer help to the participants, another perspective is tried (i.e., the process returns to Stage 3).

An example helps illustrate key parts of the method. In this case, Symon and Clegg (1991) studied the implementation of a Computer Aided Design Computer Aided Manufacturing (CAD/CAM) system in a large aerospace company in the United Kingdom. The study was undertaken between 1988 and 1990, with a follow-up in 1996. After a period of intensive data gathering (Stages 1 and 2), the researchers argued that the company was utilizing an inappropriate (and very limited) view (or relevant system) of the implementation process, and offered a different one (Stage 3). The old relevant system saw the process as a straightforward technology implementation project, whereas the new one emphasized organizational change. A new conceptual model was generated, one required to meet the needs of this new view of the CAD/CAM implementation (Stage 4). The new model stressed the need to resource, plan, participate, educate, support, and evaluate. These needs were then matched against the actual implementation activities, and this helped identify some major gaps in practice. This led to the generation of a series of recommendations for action (Stage 5). After discussion (Stage 6), some of these were implemented by the company (Stage 7).

The researchers continued their work and further evaluated the impact of the CAD/CAM as it became more widespread in use, again matching practice against the conceptual model (from Stage 4). Further gaps were identified, and a second set of recommendations for action was made to the company (Stage 5). Again, some of these were agreed upon and implemented (Stages 6 and 7).

The follow-up study 6 years later revealed that the CAD/CAM implementation had been a success insofar as the technology was fully operational, widely used, and well liked. It was meeting most of its initial objectives. Various actions instigated by the researchers

(such as the formation of a CAD/CAM User Group) continued to function effectively. However, the CAD/CAM implementation had not achieved the level of integration between design and manufacturing that had been sought, in large part because the necessary structural changes between these two powerful functions had not been attempted.

As can be seen from this example, SSA has marked similarities to ACTION RESEARCH. In this case, we can see SSA as a particular instance of action research. However, action research includes a range of different approaches (e.g., survey feedback). Furthermore, some SSA change programs are more concerned with action than they are with research, and hence should not be labeled strictly as exemplars of action research.

SSA has a number of essential characteristics (Clegg & Walsh, 1998). It is highly participative, and it provides some structure and organization to the process of analyzing problems and planning change in complex systems. At various stages in the process, it actively promotes innovative thinking, and it requires careful analysis and logical modeling. With use, the method can be internalized, and, in our view, it can be a very useful addition to the skill set of people involved in research and development.

—Chris Clegg and Susan Walsh

REFERENCES

- Checkland, P. B. (1981). *Systems thinking, systems practice*. Chichester, UK: Wiley.
- Clegg, C. W., & Walsh, S. (1998). Soft systems analysis. In G. Symon & C. M. Cassell (Eds.), *Qualitative methods and analysis in organisational research: A practical guide*. London: Sage.
- Naughton, J. (1984). *Soft systems analysis: An introductory guide*. Milton Keynes, UK: Open University Press.
- Symon, G. J., & Clegg, C. W. (1991). Technology-led change: A study of the implementation of CAD/CAM. *Journal of Occupational Psychology*, 64, 273–290.

SOLOMON FOUR-GROUP DESIGN

The Solomon four-group design (Solomon, 1949) is a method used to assess the plausibility of PRETEST SENSITIZATION effects, that is, whether the mere act

of taking a pretest influences scores on subsequent administrations of the test. On a vocabulary test, for example, participants who were given a pretest might then decide to search a dictionary for a few unfamiliar words from that test. At posttest, they might then score better on the vocabulary tests compared to how they would have scored without the pretest. META-ANALYTIC results suggest that pretest sensitization does occur, although it is more prevalent for some kinds of measures than others, and is dependent on the length of time between pretest and posttest (Willson & Putnam, 1982).

In the Solomon four-group design, the researcher randomly assigns participants to one of four cells constructed from two fully crossed factors—treatment (e.g., TREATMENT or CONTROL GROUP) and whether or not a pretest is administered. Thus, the four cells are (a) treatment with pretest, (b) treatment without pretest, (c) control with pretest, and (d) control without pretest. The analysis then tests for whether treatment works better than control, whether groups receiving a pretest score differently from groups not receiving a pretest, and whether pretesting interacts with treatment effects.

The Solomon four-group design provides information relevant to both INTERNAL and EXTERNAL VALIDITY. Regarding internal validity, for example, many single-group QUASI-EXPERIMENTS use both pretests and posttests. If a pretest sensitization effect occurs, it could be mistaken for a treatment effect, so that scores could change from pretest to posttest even if the treatment had no effect at all. In randomized EXPERIMENTS, however, pretest sensitization effects do not affect internal validity because they are randomized over conditions, but those effects can be relevant to external validity if the size of the effect is dependent on whether or not a pretest is administered.

A computer literature search found 102 studies that reported using the design or a modified version of it, more than half of which (55) were published within the past 10 years. Many of those found no pretesting effect (e.g., Haight, Michel, & Hendrix, 1998). Dukes, Ullman, and Stein (1995) did find a pretesting effect on one of four DEPENDENT VARIABLES they used, resistance to peer pressure, suggesting potential lack of generalizability of the treatment effect when comparing groups with and without a pretest.

—William R. Shadish and M. H. Clark

REFERENCES

- American Psychological Association. (2002, October 28). *PsycINFO database records* [electronic database]. Washington, DC: Author.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: RandMcNally.
- Dukes, R. L., Ullman, J. B., & Stein, J. A. (1995). An evaluation of D.A.R.E. (Drug Abuse Resistance Education), using a Solomon four-group design with latent variables. *Evaluation Review*, 19(4), 409–435.
- Haight, B. K., Michel, Y., & Hendrix, S. (1998). Life review: Preventing despair in newly relocated nursing home residents: Short- and long-term effects. *International Journal of Aging and Human Development*, 47(2), 119–142.
- Michel, Y., & Haight, B. K. (1996). Using the Solomon four design. *Nursing Research*, 45(6), 367–369.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, 46, 137–150.
- Walton Braver, M., & Braver, S. L. (1996). Statistical treatment of the Solomon four-group design: A meta-analytic approach. *Psychological Bulletin*, 104(1), 150–154.
- Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal*, 19, 249–258.

SORTING

In the method of sorting, respondents are presented with a set of objects and asked to divide the objects into a number of groups. As a method of data collection, sorting assumes that the researcher has chosen a specific area or domain and defined it by a list of instances (consisting of a word, phrase, photograph, picture, item, or actual object). These objects usually have been elicited previously by procedures such as free-listing or taxonomy construction. The respondent is presented with these objects, usually on a card, and asked to sort them into separate groups or piles (hence, pile-sorting) so that similar objects are put in the same group. The groups thus represent distinct categories or classes of the domain and together constitute a nominal scale.

The method of sorting is used especially in linguistics (to study the process of category formation) and in the study of perceptions of social and cultural worlds in order to see what significant categories and divisions

are used by participants. Sorting has the advantage that it can be used with large numbers of objects, and it is a task that respondents often report that they enjoyed doing.

VARIATIONS IN BASIC SORTING

In the basic form (free-sorting), any number of groups, of any size, may be formed. Restrictions imposed on the number and kind of groups define the variants of sorting:

- *Fixed sorting* occurs when there is a prespecified number of groups.
- *Graded sorting* occurs when groups are fixed in number and also ordered, as in a LIKERT SCALE.
- *Q-sort* occurs when, in addition to being fixed and ordered, the objects are required to follow a specified normal distribution across the groups.

Once the groups are formed, the respondent is invited to name and/or describe them, thus providing valuable interpretative data about the semantics of the categories.

More complex types of sorting include the following:

- *Multiple or successive sortings* (Boster, 1994), where respondents perform several sortings on the objects according to the same or different criteria
- *“Finer/coarser” sorting* (or GPA: Group-, Partition-, Additive-sorting) (Bimler & Kirkland, 1999), when the categories of the final sorting are (a) merged to form a grosser sorting, and (b) divided down to form a more detailed one. A very general form of sorting is the *agglomerative hierarchical sorting* (or *hierarchical clustering method*) (see Coxon, 1999), where respondents begin with each object in a single group (the “Splitter” partition) and form a set of successively more general sortings until all are in a single group (the “Lumper” partition)
- *Overlapping sorting*, where respondents may be permitted to allocate an object to more than one group, but in that case, conventional statistical methods cannot be used because the data no longer form a partition

ANALYSIS OF SORTING DATA

Once collected, sorting data can be analyzed in terms of the similarity of the *objects*, using co-occurrence measures (Burton, 1975) to measure how frequently a pair of objects occurs together in the respondents' groups, or the similarity of the *sortings* using minimum-move partition measures (Arabie & Boorman, 1973). These measures are represented using MULTIDIMENSIONAL SCALING and hierarchical clustering methods, and Takane (1981) has developed joint and three-way scaling models for these data.

—Anthony P. M. Coxon

REFERENCES

- Arabie, P., & Boorman, S. A. (1973). Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology*, 10, 148–203.
- Bimler, D., & Kirkland, J. (1999). Capturing images in a net: Perceptual modeling of product descriptors using sorting data. *Marketing Bulletin*, 10, 11–23.
- Boster, J. S. (1994). The successive pile sort. *Cultural Anthropology Methods*, 6(2), 7–8.
- Burton, M. L. (1975). Dissimilarity measures for unconstrained sorting data. *Multivariate Behavioral Research*, 10, 409–424.
- Coxon, A. P. M. (1999). *Sorting data: Collection and analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–127). Thousand Oaks, CA: Sage.
- Takane, Y. (1981). Multidimensional scaling of sorting data. In Y. P. Chaubey & T. D. Dwivedi (Eds.), *Topics in applied statistics*. Montreal: Concordia University Press.

SPARSE TABLE

A sparse table is a cross-classification of observations by two or more discrete variables that has many cells with small or zero frequencies. Sparse CONTINGENCY TABLES occur most often when the total number of observations is small relative to the number of cells. For example, consider a table with $N = 4$ cells and $n = 12$ observations. If the observations are spread evenly over the four cells, then the maximum possible frequency is small (i.e., $n/N = 3$). Furthermore, if the occurrence of observations in one of the four cells is a rare event, then a large sample would be needed to obtain observations in this cell. As a second example, consider a cross-classification of four variables, each

Table 1 Example of a Two-Way Table With Sampling Zeros

		<i>Genetic Testing: More Harm or Good?</i>			<i>Total</i>
		<i>More Good</i>	<i>More Harm</i>	<i>It Depends</i>	
<i>How Much</i>	<i>Great Deal</i>	54	10	0	64
<i>Know About</i>	<i>Not Very Much</i>	170	88	0	258
<i>Genetic Tests?</i>	<i>Nothing at All</i>	17	14	0	31
<i>Total</i>		241	112	0	353

with seven categories, which has $N = 7^4 = 2,401$ cells. A sample size considerably larger than 2,401 is required to ensure that all cells contain nonzero frequencies, and a much larger sample size is needed to ensure that all frequencies are large enough for statistical tests and modeling.

Sparseness invalidates standard statistic hypothesis tests, such as chi-square tests of independence or model goodness-of-fit. The justification for comparing test statistics (e.g., PEARSON'S CHI-SQUARE STATISTIC or the LIKELIHOOD RATIO STATISTIC) to a chi-square distribution depends critically on having "large" samples, where large means having EXPECTED VALUES for cells that are greater than or equal to 5. Without large samples, the probability distribution with which test statistics should be compared is unknown. Possible solutions to this problem include using an alternative statistic, performing exact tests, or approximating the probability distribution of test statistics by resampling or Monte Carlo methods.

Sparse tables often contain zero frequencies, which can cause estimation problems, including biased descriptive statistics (e.g., ODDS RATIO), the estimation of LOG-LINEAR MODEL parameters, and difficulties for computational algorithms that fit models to data. Whether an estimation problem exists depends on the pattern of zero frequencies in the data and the particular model being estimated. Parameters cannot be estimated when there are zeros in the corresponding margins. For example, Table 1 consists of the cross-classification of responses to the following two questions and possible responses from the 1996 General Social Survey (<http://www.icpsr.umich.edu:8080/GSS/homepage.htm>): (a) "Based on what you know, do you think genetic testing will do more good than harm?" with possible responses of "more good than harm," "more harm than good," and "it depends"; and (b) "How

much would you say that you have heard or read about genetic screening?" with possible responses of "a great deal," "not very much," and "nothing at all." None of the respondents answered "It depends" to the first question. For the independence log-linear model, a parameter for the marginal effect for "It depends" cannot be estimated. The information needed to estimate this parameter is the column marginal value, which has no observations.

In larger data sets, detecting a problem is more difficult. Signs of an estimation problem are extremely large estimates of standard errors or error messages on computer output such as "failure to converge." Solutions to estimation problems include adding parameters to a model so that the empty cells causing a problem are fit perfectly, adding a very small value to all cells of the table (e.g., 1×10^{-8}), or using an alternative estimation method. When empty cells exist because they are theoretically impossible (e.g., a man with menstrual problems), an extra parameter should be added to log-linear models for each empty cell. Adding parameters removes the cell from the analysis.

—Carolyn J. Anderson

REFERENCES

- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.
 Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

SPATIAL REGRESSION

Spatial REGRESSION analysis consists of specialized methods to handle complications that may arise when applying regression methods to spatial (geographic) data or to models of spatial interaction. There are

two broad classes of spatial effects. One is referred to as *spatial HETEROGENEITY* and occurs when the parameters of the model (regression coefficients, error variance); the functional form itself; or both are not stable across observations. This is a special case of structural instability and can be handled by means of the usual methods (modeling HETEROSKEDASTICITY, varying coefficients, random coefficients). It becomes *spatial* heterogeneity when the variability shows a spatial structure, which then can be exploited to model the heterogeneity more efficiently. For example, in the analysis of determinants of criminal violence, the southern states of the United States have been found to have different regression coefficients from the rest of the country. Because those states are geographically defined, this is referred to as a spatial regime, a special form of spatial heterogeneity.

The other form of spatial effects is referred to as *spatial dependence* (or *spatial AUTOCORRELATION*) and occurs when the error terms in the regression model are spatially correlated or when spatial interaction terms are included as explanatory variables in the regression specification. Spatial error autocorrelation is a special case of an error covariance matrix with nonzero off-diagonal elements. It is “spatial” when the pattern of nonzero elements follows from a model of spatial interaction. Typically, this involves a notion that neighbors are more correlated than locations further apart. For example, in real estate analysis, the sales price of a house is commonly considered to be determined in part by neighborhood effects. However, the latter cannot be quantified and therefore are incorporated in the regression error term. Because neighboring houses will be exposed to the same neighborhood effect, the regression errors will be spatially correlated.

The formal specification of the pattern of spatial correlation requires additional structure. There are two general ways in which this is pursued. In one, referred to as *geostatistical modeling*, the correlation between pairs of observations is expressed as a decreasing function of the distance that separates them. This is formally incorporated into the semivariogram in the field of geostatistics. In the other, referred to as *lattice* or *regional modeling*, the correlation follows from a spatial STOCHASTIC process, where the value of a random variable at a given location is expressed as a function of the other random variables at other locations through a spatial weights matrix. The spatial weights matrix is a positive matrix in which neighbors have a nonzero value (by convention, the diagonal elements are set to

zero). The definition of neighbors is perfectly general and need not be based on geographic considerations (e.g., social networks).

Apart from error correlation, spatial interaction can also be incorporated in the form of a spatially lagged dependent variable (or spatial lag), which is essentially a weighted average of the values for the dependent variable at neighboring locations. The resulting mixed regressive-spatial autoregressive model is the most commonly used specification for spatial externalities and spatial interaction in a cross-sectional setting. For examples, in models used in the study of local public finance, the tax rates set in one district may be expressed as a function of the rates in the neighboring districts (a spatially lagged dependent variable).

Ignoring spatial effects affects the properties of the ORDINARY LEAST SQUARES (OLS) ESTIMATOR. Ignoring a spatial lag renders OLS inconsistent, whereas ignoring a spatially correlated error term results in inefficient (but still unbiased) estimates. Specialized estimation methods have been developed, based on the maximum likelihood principle, instrumental variables, and general method of moments. Diagnostics for spatial autocorrelation are based on Moran's I autocorrelation statistic and applications of the Lagrange multiplier principle. Recent extensions include the incorporation of spatial correlation in models for latent variables and for count data, in the form of hierarchical specifications. Estimation of such models typically requires the application of SIMULATION estimators, such as the GIBBS SAMPLER.

—Luc Anselin

REFERENCES

- Anselin, L. (1988). *Spatial econometrics*. Boston: Kluwer Academic.
- Anselin, L., & Bera, A. (1998). Spatial dependence in linear regression models, with an introduction to spatial econometrics. In A. Ullah & D. Giles (Eds.), *Handbook of applied economic statistics* (pp. 237–289). New York: Marcel Dekker.
- Cressie, N. (1993). *Statistics for spatial data*. New York: Wiley.

SPEARMAN CORRELATION COEFFICIENT

A special case of the PEARSON'S CORRELATION COEFFICIENT, the Spearman correlation coefficient

Table 1 Ranks of 20 Pianists on Technical Difficulty and Artistic Expression by a Judge

<i>Technical Difficulty (X)</i>	<i>Artistic Expression (Y)</i>	X^2	Y^2	XY	$d^2 = (X - Y)^2$
1	2	1	4	2	1
6	4	36	16	24	4
3	6	9	36	18	9
4	3	16	9	12	1
5	1	25	1	5	16
2	5	4	25	10	9
7	7	49	49	49	0
8	9	64	81	72	1
9	12	81	144	108	9
17	10	289	100	170	49
11	8	121	64	88	9
12	11	144	121	132	1
13	13	169	169	169	0
14	15	196	225	210	1
15	14	225	196	210	1
16	20	256	400	320	16
10	18	100	324	180	64
18	16	324	256	288	4
19	19	361	361	361	0
20	17	400	289	340	9
$\sum X = 210$	$\sum Y = 210$	$\sum X^2 = 2,870$	$\sum Y^2 = 2,870$	$\sum XY = 2,768$	$\sum d^2 = 204$

$$r = \frac{2768 - \frac{210 \times 210}{20}}{\sqrt{2870 - \frac{210^2}{20}} \sqrt{2870 - \frac{210^2}{20}}} = 0.85$$

$$r_s = 1 - \frac{6 \sum d^2}{n^3 - n} = 1 - \frac{6 \times 204}{20^3 - 20} = 0.85$$

NOTE: The Pearson correlation coefficient (r) is applied and shown on the left-hand side. Note that both r and r_s reach the same conclusion.

(r_s) is one of the most widely used nonparametric correlation coefficients and gauges the relationship between two variables measured at the ordinal level. Suppose creativity (denoted X) and productivity (denoted Y) of n subordinates are evaluated by a manager. The manager assigns one of the ranks (i.e., $1, 2, \dots, n$) to all employees based on their creativity and productivity, respectively. Assuming there are no tied ranks on both variables, the relationship between creativity and productivity can be assessed by the Spearman correlation coefficient,

$$r_s = 1 - \frac{6 \sum d^2}{n^3 - n}$$

(where d is the difference between X and Y), which can actually be simplified from the formula for Pearson's correlation coefficient (Chen & Popovich, 2002).

A careful examination of the simplified formula reveals that the greater the difference between two variables across n subjects, the greater the value of $\sum d^2$ and the smaller the correlation coefficient. A fictitious example of 20 pianists' competition results is provided in Table 1. Each pianist is evaluated based on technical difficulty and artistic expression. Suppose a judge assigns ranks of 1 to 20 to each contestant on both performance categories, where rank 1 represents the highest performance. Assuming there are no tied ranks, the relationship between the two performance categories calculated by either the Spearman correlation coefficient or the Pearson correlation coefficient is 0.85. It should be emphasized that the above simplified formula cannot correctly calculate the Spearman correlation coefficient if tied ranks exist. The easiest way to calculate the

Table 2 Relationship Between Numbers of Annual Patent Applications and Quarterly Sales (in Millions)

<i>Organizations</i>	<i>Patent Application (X)</i>	<i>Quarterly Sales (Y)</i>	<i>Rank of X (RX)</i>	<i>Rank of Y (RY)</i>
A	20	5.60	5.0	6.5
B	9	1.10	2.0	1.0
C	50	6.60	9.0	9.0
D	33	7.20	8.0	10.0
E	17	4.30	3.0	4.0
F	25	5.60	6.0	6.5
G	79	6.30	10.0	8.0
H	5	2.90	1.0	2.0
I	27	4.90	7.0	5.0
J	19	3.70	4.0	3.0
Correlation	Pearson correlation coefficient based on the untransformed data (<i>X</i> and <i>Y</i>) is 0.68.		Spearman correlation coefficient based on the transformed data (<i>RX</i> and <i>RY</i>) is 0.89.	

NOTE: Organizations A and F are tied on quarterly sales. An average of rank 6 and rank 7, 6.5, is assigned to both organizations.

Spearman correlation coefficient is to apply the Pearson correlation coefficient whenever either tied or nontied ranks exist.

If the Spearman correlation coefficient is applied to variables measured on INTERVAL or RATIO scales, ranks of the variables need to be artificially assigned first. Numbers of annual patent applications (*X*) and quarterly sales (*Y*) of 10 organizations are presented in Table 2. Both variables are measured as ratio scales. Prior to applying the Spearman correlation coefficient, values of both variables should be transformed into ranks, where rank 10 represents the highest performance. Following this transformation, the Spearman correlation coefficient can be obtained with the same procedure described earlier. Note that tied ranks exist for the quarterly sales variable (Organizations A and F). Hence, the formula

$$r_s = 1 - \frac{6 \sum d^2}{n^3 - n}$$

is no longer appropriate, and the Spearman correlation should be obtained by the Pearson correlation coefficient formula. The magnitude of the Pearson correlation coefficient calculated based on the original values tends to be greater than that of the Spearman correlation coefficient calculated based on the transformed ranks, although the example provided here shows the exception.

The null hypothesis test of $\rho_s = 0$ can be conducted by

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}},$$

$df = n - 2$ when sample size is $19 \leq n < 30$. However, the test of $\rho_s = 0$ can be conducted more accurately by $z = r_s \sqrt{n-1}$ if $n > 35$ (Kendall & Gibbons, 1990). Exact critical values should be used when n is small for both formulas because the sampling distribution of r_s with a small sample size is not distributed as either the t or z distribution. When n is large, there is no apparent difference between both tests (Siegel & Castellan, 1988).

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

- Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage.
- Kendall, M., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). New York: Oxford University Press.
- Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

SPEARMAN-BROWN FORMULA

Developed independently by Charles Spearman (1910) and William Brown (1910), this formula

estimates the RELIABILITY of multi-item measures or INDEXES, and provides an important foundation for the statistical theory of measurement. The Spearman-Brown formula also calculates expected increases in reliability if more INDICATORS of the same trait are available. In both capacities, the Spearman-Brown formula is used frequently in the social sciences, especially in sociology, psychology, and education, where multi-item indexes are commonly employed.

To illustrate, following Brown (1910), let x_1 and x_2 be two measurements of a particular characteristic x , together comprising a “test” in which x_1 and x_2 are given equal weight. x'_1 and x'_2 are two other measurements of the same trait and, in combination, provide a parallel test of characteristic x . Suppose that all of the measurements are equally reliable, or that

$$\sigma_{x_1} = \sigma_{x_2} = \sigma_{x'_1} = \sigma_{x'_2} = \sigma_x,$$

where σ_{x_1} is the standard deviation of x_1 , and that their relationships to each other are comparable,

$$\sigma_{x_1x'_1} = \sigma_{x_1x'_2} = \sigma_{x_2x'_1} = \sigma_{x_2x'_2} = n\sigma_x^2r,$$

where $\sigma_{x_1x'_1}$ is the covariance of x_1 and x'_1 , r is a measure of reliability of each test, and n is the number of tests (in this case, $n = 2$). In practice, as Nunnally and Bernstein (1994) suggest, r might be the average CORRELATION across the group of observed measures (in this example, between x_1 and x_2) comprising a test. Then, the reliability of the indicator is estimated using an adjusted correlation coefficient, following Brown (1910):

$$\begin{aligned} r_{SB} &= \frac{\sigma_{(x_1+x_2)(x'_1+x'_2)}}{n\sigma_{(x_1+x_2)}\sigma_{(x'_1+x'_2)}} \\ &= \frac{\sigma_{x_1x'_1} + \sigma_{x_1x'_2} + \sigma_{x_2x'_1} + \sigma_{x_2x'_2}}{n(\sigma_{x_1}^2 + \sigma_{x_2}^2 + 2r\sigma_{x_1}\sigma_{x_2})} \\ &= \frac{4n\sigma_x^2r}{n(2\sigma_x^2 + 2r\sigma_x^2)} \\ &= \frac{4n\sigma_x^2r}{2n\sigma_x^2(1+r)} \\ &= \frac{2r}{1+r}. \end{aligned}$$

More generally, the Spearman-Brown coefficient may be calculated

$$r_{SB} = \frac{nr}{1 + (n-1)r},$$

where, as above, r is a measure of reliability (or average correlation among parallel tests) and n is the number of parallel tests.

To demonstrate, consider the following example: William H. Sewell (1942) was interested in developing a measure of the socioeconomic status of farm families. From an initial list of 123 items, Sewell constructed an index of 36 indicators he believed to measure a family's status. Then, dividing the items according to their numbering (even and odd) to construct two “half-tests,” Sewell estimated the correlation between them to be 0.83. Using the Spearman-Brown formula, Sewell adjusted this estimate of the correlation between the half-tests to generate an estimate of the reliability of his measure that takes the full test length into account:

$$\begin{aligned} r_{SB} &= \frac{nr}{1 + (n-1)r} \\ &= \frac{2(0.83)}{1 + (2-1)(0.83)} \\ &= 0.91. \end{aligned}$$

Because the Spearman-Brown coefficient is bounded by 0 and 1, where 1 indicates the highest possible level of reliability, Sewell was well satisfied by this result. If, however, Sewell was interested in increasing the reliability of his measure, the Spearman-Brown formula predicts gains in reliability that result from doubling the number of indicators, such that each half-test would now comprise one of four parallel tests (provided the original assumptions hold):

$$\begin{aligned} r_{SB} &= \frac{nr}{1 + (n-1)r} \\ &= \frac{4(0.83)}{1 + (4-1)(0.83)} \\ &= 0.95. \end{aligned}$$

Similarly, the Spearman-Brown “prophecy” formula can be used to estimate the number of indicators needed to achieve a desired level of reliability in a measure.

Analysts should note two assumptions made in the derivation of the Spearman-Brown formula: First, the expected values and variances of each measure are assumed to be equal across measures. Second, as stated above, covariances among all pairs of measures are assumed to be equivalent. In other words, the measures, on average, should generate the same mean scores, should be equally precise, and should

be equally related to each other. These assumptions, describing “classical parallelism” or an “exchangeable covariance structure” (Li, Rosenthal, & Rubin, 1996), are important because they allow analysts to estimate the reliability of an index with an adjusted measure of the correlation among its components. It seems, however, that these conditions rarely occur in the social sciences. Recognizing this, some researchers have provided generalized versions of the Spearman-Brown formula that readily incorporate measures with different levels of precision: In particular, Lee Cronbach, and G. Frederick Kuder and Marion Richardson, have built upon the important foundation laid by Spearman and Brown in the statistical theory of measurement; see CRONBACH’S ALPHA in this volume.

—Karen J. Long

REFERENCES

- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 12, 296–322.
- Li, H., Rosenthal, R., & Rubin, D. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, 1(1), 98–107.
- Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Sewell, W. H. (1942). The development of a sociometric scale. *Sociometry*, 5(3), 279–297.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 12, 271–295.

SPECIFICATION

Model specification refers to the description of the process by which the DEPENDENT VARIABLE is generated by the INDEPENDENT VARIABLES. Thus, it encompasses the choice of independent (and dependent) variables, as well as the functional form connecting the independent variables to the dependent variable. Specification can also include any ASSUMPTIONS about the STOCHASTIC component of the MODEL. Thus, specification occurs before ESTIMATION. A simple example of model specification with which most people are familiar is a model that assumes that some variable Y is a LINEAR FUNCTION of a RANDOM variable X and a stochastic disturbance term ε :

$$Y = \alpha + \beta X + \varepsilon. \quad (1)$$

There are many alternative specifications that would relate Y to X , such as

$$Y = \alpha + \beta X^2 + \varepsilon, \text{ or} \quad (2)$$

$$Y = \alpha + \beta X^3 + \varepsilon. \quad (3)$$

Once a model is specified, an estimation technique is merely a tool for determining the values of the PARAMETERS in the model that best fits the observed data to the specification. However, without correct model specification, any estimation exercise is merely curve-fitting without any ability to make valid STATISTICAL INFERENCES. To make inferences from the observed SAMPLE of data to the POPULATION, we are required to assume that all values of Y in the population are generated via the same process that generates the values in our observed sample. In other words, we must assume that the specification is correct. In the well-known GAUSS-MARKOV THEOREM, this is generally stated as “the model is correctly specified.”

Model specification should be based on some relevant THEORY. If the estimated parameters of the model are STATISTICALLY SIGNIFICANT, it may be regarded as confirmation of the theory. But in fact, it is technically a test of the model as specified versus the null model—a model that assumes there is no relationship between the dependent variable and the independent variables of the model. If we estimated any of the three models above and were able to reject the NULL HYPOTHESIS that $\beta = 0$ at the 95% level, we still could not say the specification we estimated was correct and the other two proposed specifications were incorrect.

Although infinitely many forms of *specification error* are possible, there are several common forms of specification error to beware of. Most ubiquitous is the problem of OMITTED VARIABLES. If the model specified omits an independent variable that is, in fact, part of the true model generating the data, then the model is misspecified. The consequences of this form of specification error are severe. Even in the simple case of ORDINARY LEAST SQUARES and a linear model, the estimated parameters of other variables may be BIASED, and we can make no valid statistical inferences.

There is no test to determine that we absolutely have the correct specification. However, we can often compare our specification to alternative specifications with statistical tests. The simplest tests involve specifications that are nested within one another. One specification is nested in another if it has identical functional form, but the variables in it are a subset

of the variables in the other specification. So, for instance, the model $Y = \alpha_0 + \beta_1 X_1 + \varepsilon$ is nested within the model $Y = \alpha_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, but the model $Y = \alpha_0 + \beta_3 X_3 + \varepsilon$ is not tested in that model. Nested models are typically tested against one another using F TESTS of linear restrictions (or CHOW TESTS) for linear models, and log-likelihood ratio tests for MAXIMUM LIKELIHOOD models. Such tests are described in almost all econometrics texts. In both cases, one estimates each of two proposed models and generates a test statistic based on the fit of each model. A classical hypothesis test can then be performed that may let one reject one model in favor of another. In the case of an F test to compare two linear specifications, one of which is nested within another, the test statistic is given by

$$F = \frac{(R_{UR}^2 - R_R^2)/m}{(1 - R_{UR}^2)/(n - k)}, \quad (4)$$

which is distributed as an F statistic with DEGREES OF FREEDOM $(m, n - k)$, where m is the number of linear restrictions, n is the number of observations, and k is the total number of right-hand-side variables.

Non-nested specification tests are generally more difficult to perform. However, there are existing tests. The Cox test allows for a statistical test between two non-nested models. The Vuong test also allows for comparisons between non-nested models.

An alternative way to select between competing model specifications is by using model selection criteria. Model selection *criteria* do not provide a statistical test to prefer one model to another. However, they do provide agreed-upon criteria to choose one model over another. They are generally based upon a trade-off between GOODNESS OF FIT (all else being equal, we would prefer a model that explains more variance to one that explains less) and parsimony (almost all selection criteria penalize specifications for inclusion of additional variables). Commonly used criteria include the Akaike Information Criterion (AIC), Schwartz Criteria (SC), and Bayes Information Criteria (BIC). The AIC is simple to compute for linear models:

$$\text{AIC}(k) = e^{2k/n} \frac{\sum e_i^2}{n}. \quad (5)$$

Frequently, this is given in terms of its log:

$$\ln \text{AIC}(k) = \frac{2k}{n} + \ln \left(\frac{\sum e_i^2}{n} \right). \quad (6)$$

Those criteria are sensible, but ultimately involve arbitrary trade-offs. In the absence of a true statistical test, the researcher is not able to make any sort of probabilistic statement of how confident he or she is that he or she made the right choice among the models.

The assumption that we have the right specification is one of the most controversial in ECONOMETRICS. It led to Leamer's (1983) famous article "Let's Take the Con out of Econometrics," where he suggested dropping the assumption and instead examining virtually all sensible combinations of right-hand-side variables and reporting the bounds on the parameters. Because such bounds are generally large, this is not a suggestion that practitioners have adopted. But it does suggest that researchers test specifications for robustness to perturbations.

—Jonathan Nagler

REFERENCE

- Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, 73(1), 31–43.

SPECTRAL ANALYSIS

Spectral analysis is one method of identifying the cyclical components of TIME-SERIES data. Cycles are periodic events typically represented as a waveform of the trigonometric sine or cosine function, or sinusoid, represented graphically and mathematically by Figure 1:

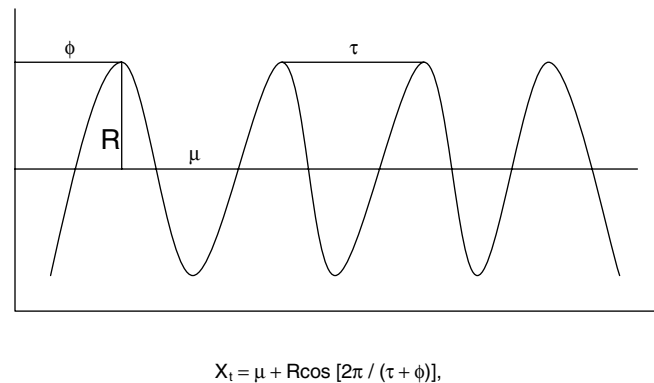


Figure 1 Spectral Analysis

where τ represents the period of the waveform, or the distance from one peak of the cycle to another; μ

represents the mean of the highest and lowest points of the waveform; ϕ represents the initial phase of the wave; and R represents the amplitude, or distance from peak to trough, of the waveform.

The sinusoid is an important component of cyclical analysis because it is used to MODEL the shape of those cycles. Where complex waveforms exist, researchers and statistical packages may model the cycles using several sinusoidal components. Although it is still possible to model the data if cycles deviate substantially from the sinusoid, such methods are mathematically complex and less tractable than the model presented here.

The cyclical nature of time-series data can be approximated once time TRENDS are removed from the data by partitioning the residual variance in a time series that is not due to a linear or CURVILINEAR trend in the data. The total SUM OF SQUARES is divided into a set of $N/2$ sum of squares components, each with 2 DEGREES OF FREEDOM. The estimates derived from the calculations are transformed into the sum of squares accounted for by each periodic component, to create intensity estimates.

In order to shrink the wide CONFIDENCE INTERVALS often associated with intensity estimates, spectral analysis smoothes the estimates to create a power spectrum. SMOOTHING procedures compute weighted averages of intensity estimates within a few neighboring frequencies. Variations in the number of neighboring frequencies (or width) included in the weighted average and variations in the weights applied, resulting from the use of different smoothing techniques, can produce more reliable estimates of the cyclical nature of the data. For instance, Daniell (equal weight) smoothing techniques with a width (or M) of 3 replace the intensity estimates for frequency i , with the mean for three ordinates, including $i - 1$, i , and $i + 1$. Equal weights (or $1/M$) are applied to the new estimates, $(1/3) \times \text{intensity at frequency } (i - 1) + (1/3) \times \text{intensity at frequency } i + (1/3) \times \text{intensity at frequency } (i + 1)$, to achieve a smoothed estimate of cycles. It is worth noting that the size of confidence intervals decreases as the number of neighboring frequencies increases. However, excessive smoothing can distort the data and make it difficult to make clear judgments about the lengths of cycles in the data (or τ).

Once the power spectrum is calculated, it is divided by the total amount of power or variance, producing a spectral density function. The results are typically

plotted with the spectral density on the y -axis and the frequency ($1/\tau$) on the x -axis. Data that exhibit a 4-year cycle would register a frequency of $1/4$ or .25 at its peak, which would be visible on a smoothed graph. Statistical significance is determined through the estimation of confidence intervals based on CHI-SQUARE DISTRIBUTIONS,

$$\text{Lower bound} = [edf^*s(f_1)]/X^2.99$$

$$\text{Upper bound} = [edf^*s(f_1)]/X^2.01,$$

where edf is the product of the original degrees of freedom prior to smoothing (2) and the width of the smoothing technique, and $s(f_1)$ denotes the spectral estimate for a given frequency. If confidence intervals overlap with the mean level of power (or mean of spectral estimates), then we might conclude that the peaks are not significantly different from what one might expect by chance. Alternatively, one could conduct a test of SIGNIFICANCE for unsmoothed peaks by dividing the largest sum of squares, or intensity estimate (sum of squares_{largest peak}), by the sum of the sum of squares total ($\sum$ sum of squares_{total}).

—Erika Moreno

REFERENCES

- Granger, C. W. J. (1964). *Spectral analysis of economic time series*. Princeton, NJ: Princeton University Press.
 Warner, R. M. (1998). *Spectral analysis of time-series data*. New York: Guilford.

SPHERICITY ASSUMPTION

The sphericity assumption is necessary for statistical inference using univariate (or mixed-model) ANALYSIS OF VARIANCE on a WITHIN-SUBJECTS DESIGN. Informally stated, the sphericity assumption assumes that the VARIANCE of any linear combination of the repeated measures is identical. In particular, the differences between treatment levels have equal variance. If the assumption is not satisfied, it is necessary to either employ statistical corrections or else use MULTI-VARIATE ANALYSIS OF VARIANCES AND COVARIANCES.

An example clarifies the assumption. Suppose three individuals are exposed to three treatments producing the following data (and differences) (see Table 1):

Table 1

<i>Treat 1</i>	<i>Treat 2</i>	<i>Treat 3</i>	<i>1-2</i>	<i>1-3</i>	<i>2-3</i>
5	7	2	-2	3	5
3	7	4	-4	1	3
6	8	6	-2	0	2
Variance			1.33	2.33	2.33

In order to be justified in treating the data as a two-factor design (i.e., TWO-WAY ANOVA), the variances of the treatment differences must be identical. As the last three columns show, the sphericity assumption is not satisfied for the hypothetical example above because the variances of the differences are not identical. Note that if there are only two treatments, the sphericity assumption is satisfied by definition.

Incorrectly ignoring violations of the sphericity assumption results in a loss of statistical POWER OF A TEST. The actual probability of a TYPE I ERROR when the sphericity assumption is not satisfied is higher than standard calculations would imply. In other words, researchers risk making inferences that are unsupported by the data because the null hypothesis of no treatment differences will be rejected more often than it should (but by an unknown amount). Note that if the null hypothesis of no difference is not rejected, no correction is necessary because the nonadjusted test statistic is biased against this possibility.

If the sphericity assumption is not satisfied, the researcher has two options. First, one could try to adjust the test statistic to account for the violation. A number of adjustments are available and detailed elsewhere (e.g., Maxwell & Delaney, 1990). Alternatively, multivariate analysis of variances and covariances could be employed. Most recommend the use of multivariate methods unless sample sizes are very small.

—Joshua D. Clinton

REFERENCES

- Bogartz, R. S. (1994). *An introduction to the analysis of variance*. Westport, CT: Praeger.
- Field, A. (1998). A bluffer's guide to . . . sphericity. *British Psychological Society: Mathematical, Statistical & Computing Newsletter*, 6, 13–22.
- Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

SPLINE REGRESSION

Spline regressions are REGRESSION line segments that are joined together at special join points known as spline knots. By definition, there are no abrupt breaks in a splined line. However, at their knots, splines may suddenly change direction (i.e., SLOPE), or, for curved lines, there may be a more subtle, but just as sudden, change in the rate at which the slope is changing.

Splines are used instead of QUADRATIC, cubic, or higher-order POLYNOMIAL regression curves because splines are more flexible and are less prone to MULTICOLLINEARITY, which is the bane of polynomial regression. Spline regressions produce smoother lines than INTERRUPTED regression, which allows breaks, or jumps, in the line itself.

Splines are useful when a change in policy affects a line's slope without causing a break or gap in the line at the moment the policy change is introduced. For example, announcing an energy tax credit for more home insulation can reduce the slope of the energy consumption curve without causing a sudden drop in energy consumption. Tax schedules typically have tax brackets where the line representing the amount of tax paid as income rises does not change abruptly when passing from one tax bracket to another but instead changes slope at these spline knots.

Piecewise linear splines appear as straight lines with kinks in them. The kinks are the changes in slope at the knot locations. Define $D = 0$ when age is less than 21, so spline $D(\text{age}-21) = 0$ for ages 0 to 21. Define $D = 1$ for over 21, so $D(\text{age}-21) = 1, 2, 3, \dots$ for ages 22, 23, 24, etc. Thus, when age passes 21, the straight line rotates to a new slope at the kink (i.e., at the spline knot location) without causing a jump or break in the line.

Spline regressions are easiest to estimate when the number of spline knots and their locations are known in advance. Announced policy changes and tax brackets would be good examples. ORDINARY LEAST SQUARES can be used to easily estimate such spline regressions.

However, sometimes, the knot locations are not known exactly. For example, people's incomes typically rise as they get older, but then, on average, begin to decline after some unspecified point. Can this

point be ESTIMATED more precisely? The location of such an unknown spline knot can be estimated using NONLINEAR least squares.

What if the number of spline knots is unknown? STEPWISE REGRESSION can be used to search for the number and location of unknown spline knots. A list of potential spline knots is created in the form of variables available for possible inclusion in a regression. A particular spline knot is invited into the regression if its location and degree of adjustment (slope, rate of change of slope, etc.) provide the greatest degree of STATISTICAL SIGNIFICANCE among all of the specified spline knot candidates.

The spline variable might be time or some CROSS-SECTIONAL variable such as a person's age or the level of production. Splines can be used to estimate three-DIMENSIONAL surfaces. For example, after controlling for the variables on the multiple listing card (bedrooms, bathrooms, school district, etc.) a three-dimensional spline regression can be used to estimate the location value of a home.

—Lawrence C. Marsh

REFERENCES

- Marsh, L. C., & Cormier, D. R. (2001). *Spline regression models*. Thousand Oaks, CA: Sage.
- Pindyck, R. S., & Rubinfeld, D. L. (1998). *Econometric models and economic forecasts* (4th ed.). New York: McGraw-Hill.

SPLIT-HALF RELIABILITY

In classical test score theory, $X = T + e$ (where X = observed score, T = true score, and e = error). Some error is constant, reducing test validity, and some is variable, leading to inconsistency of scores. Reliability is the degree to which the test is free from variable error. Formally, given that $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$, reliability is the ratio of true score variance to error score variance:

$$\sigma_t^2/\sigma_x^2 = (\sigma_x^2 - \sigma_e^2)/\sigma_x^2 \quad \text{or} \quad 1 - \sigma_e^2/\sigma_x^2.$$

Because they are necessarily a sample from the domain being measured, test items constitute one source of variable error. It can be estimated by comparing scores from two equivalent (alternate) forms of the test given at different times, but performance on the second test may be affected by the first one. Split-half reliability requires only one test administration, so participants are unaware of any comparison. It

is obtained by splitting the items into two equivalent sets and by correlating the pairs of scores (Pearson product-moment coefficient, r). Many kinds of split are possible, but some of them will not lead to equivalent sets. For example, item difficulty may vary between two sets selected at random and between two sets created from the first and second halves of the test. Scores on the two halves may also differ because of practice or fatigue, artificially lowering the split-half estimate.

Given that any split-half reliability value is not unique, at least two kinds should be calculated. There is even an algorithm for maximizing true reliability. However, the most popular technique is to group odd and even items. This overcomes fatigue and practice effects and equates the sets on difficulty if it increases throughout the test. Unfortunately, the odd/even split may not guarantee equivalence if different kinds of items are alternated, such as those requiring true/false or yes/no responses. The best method is to systematically create two equivalent sets by matching items according to content, difficulty, and kind of response (the parallel split). Means and variances of the two sets should be equal.

Because test reliability is a positive but decelerating function of the number of items, split-half r will underestimate full test reliability. Therefore, it is corrected with the SPEARMAN-BROWN prophecy formula $r' = 2r/(1 + r)$. Full test reliability can also be estimated directly with formulae such as Rulon's $r'' = 1 - \sigma_d^2/\sigma_x^2$ (where σ_d^2 is the variance of difference scores from the two parts of the test). Notably, this reflects the classic definition of reliability ($1 - \sigma_e^2/\sigma_x^2$, where $\sigma_e^2 = \sigma_d^2$). It yields the same value as the Spearman-Brown correction if $\sigma_1^2 = \sigma_2^2$ but is lower if $\sigma_1^2 \neq \sigma_2^2$.

Split-half reliability is not appropriate for tests that are highly speeded because people will score lower or higher on *both* sets, leading to a spuriously high correlation. It is also unsuitable for multidimensional tests because the different parts may be unrelated. However, it is a convenient technique for a homogeneous test. Although there are no universal standards for reliability, split-half (error from items within the test) should normally exceed .80, perhaps .85, and usually will be higher than TEST-RETEST (error from time) and ALTERNATE FORM (error from different items and time).

Example 1: A 24-item test taken by 148 people ($M = 11.20, s^2 = 20.20$) was divided into two sets

of 12 items. The means (5.38 and 5.89) and variances (7.12 and 5.57) for each set were not significantly different, indicating equivalent sets. The r between sets was .626, and the $s_d^2 = 4.81$. Using Spearman-Brown, $r' = 2(.626)/(1 + .626) = .770$. Using Rulon, $r'' = 1 - 4.81/20.20 = .762$. Both values indicate that split-half reliability is less than desirable.

Example 2: A 20-item test (5-point response scale) taken by 59 people ($M = 56.61$, $s^2 = 240.449$) was divided into odd and even items. The means (28.37 and 28.24) and variances (68.96 and 59.25) were not significantly different, indicating equivalent sets. The r between sets was .878 and $s_d^2 = 15.98$. Using Spearman-Brown, $r' = 2(.878)/(1 + .878) = .935$. Using Rulon, $r'' = 1 - 15.98/240.449 = .934$. Both values indicate that split-half reliability is very good.

—Stuart J. McKelvie

REFERENCES

- Callendar, J. C., & Osburn, H. G. (1977). A method for maximizing split-half reliability coefficients. *Educational and Psychological Measurement*, 37, 819–825.
- Cronbach, L. J. (1943). On estimates of test reliability. *Journal of Educational Psychology*, 34, 485–494.
- Cronbach, L. J. (1946). A case study of the split-half reliability coefficient. *Journal of Educational Psychology*, 37, 473–480.
- Gulliksen, H. (1950). *The theory of mental tests*. New York: Wiley.
- Lord, F. M. (1956). Sampling error due to choice of split in split-half reliability coefficients. *Journal of Experimental Education*, 24, 245–249.

SPLIT-PLOT DESIGN

A split-plot design is a research design in which units of the EXPERIMENT (e.g., consumers, parents, students) receive or are assigned to all treatments of one factor (say, Factor *B*) but only one treatment of the other factor (say, Factor *A*). In a split-plot design, Factor *A* is also called a “between-subjects” factor, whereas Factor *B* is called a “within-subjects” factor. For example, say one wishes to study the effect of Internet search strategies (Factor *A*) on college students’ information-seeking efficiency (the dependent variable) in two areas: health and history (levels of Factor *B*). One hundred freshmen at a state college are randomly selected and assigned to two conditions of Factor *A*: 50 students to the “linear search”

condition and the remaining 50 to the “nonlinear search” condition. All subjects are instructed to search for pertinent information on two topics chosen from health and history. The success of search is judged by three criteria: completeness, relevance, and speed. The following diagram illustrates the research design:

		Factor B: Topics			
		Subject	Health	History	Row Mean
Factor A: Internet Search Strategy	Treatment 1	S ₁	$\bar{Y}_{11}$	$\bar{Y}_{12}$	$\bar{Y}_{1.}$
	Linear	.			
	Search	.			
	Search	.			
	S ₅₀				
	Treatment 2	S ₅₁	$\bar{Y}_{21}$	$\bar{Y}_{22}$	$\bar{Y}_{2.}$
	NonLinear	.			
	Search	.			
	.	.			
	S ₁₀₀				
Column Mean			$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	Grand Mean = $\bar{Y}_{..}$

In the diagram and research design, both Factors *A* and *B* are fully crossed. Because of the crossed nature of both factors, this design is technically a split-plot factorial (or SPF) design. Furthermore, an equal number of subjects are assigned to each condition of Factor *A*, which makes the design a balanced SPF design. Otherwise, it is an unbalanced SPF design. Because Factor *A* has two levels (linear search strategy vs. nonlinear search strategy) and Factor *B* also has two levels (health vs. history), this design is referred to in textbooks as an SPF_{2,2} design. The number of levels of a between-subjects factor is always written before the period (.), and the number of levels of a within-subjects factor is written after the period, both in the subscript. Variations of the split-plot design are in the form of more complex designs, such as the SPF_{pq,r} design, with two between-subjects factors and one within-subjects factor; the SPF_{p,qr} design, with one between-subjects factor and two within-subjects factors; the SPF_{pq,rt} design, with two between-subjects factors and two within-subjects factors; and so on. For details on these variations, readers should consult Kirk (1995) and Maxwell and Delaney (1990).

Notice that, in the SPF design above, subjects (denoted as S in the diagram) are embedded in a treatment condition of Factor *A* (or the between-subjects

factor) throughout the study. Consequently, an SPF design is also a nested design because subjects are considered nested within levels of Factor A (or the between-subjects factor). Furthermore, an SPF design is also a block design because subjects serve as their own controls in all levels of Factor B (or the within-subjects factor); hence, they are considered “blocks” according to the rationale of randomized block designs.

Split-plot design is one of the most popular research designs in social sciences. The design originated in agricultural research, in which parcels of land are called plots. Plots are subdivided into subplots in order to investigate the effect of fertilizers, soil conditions, and treatments of this sort. E. F. Lindquist (1953) popularized the use of this design in psychology and education and renamed it a MIXED DESIGN. The term *mixed design* reflects the unique mixture of the between-subjects factor and the within-subjects factor in the design. The split-plot design is also regarded as MODEL III design (Hays, 1994) in the sense that blocks (or subjects, in the example above) are usually treated as a random factor, whereas Factor A or B , or both, are a fixed factor. The inclusion of both random and fixed factors in the same SPF design dictates the statistical model, as stated below, to be a mixed model.

$$Y_{ijk} = \mu + \alpha_j + \pi_{i(j)} + \beta_k \\ + (\alpha\beta)_{jk} + \beta\pi_{ki(j)} + \varepsilon_{ijk}$$

$$(i = 1, \dots, n; j = 1, \dots, p; \text{ and } k = 1, \dots, q),$$

where

Y_{ijk}	is the score of the i th block (also the unit of observation and analysis) in the j th level of Factor A and k th level of Factor B .
μ	is the grand mean, therefore, a constant, of the population of all observations.
α_j	is the effect of the j th treatment condition of Factor A ; algebraically, it equals the deviation of the population mean (μ_j) from the grand mean (μ). It is a constant for all observations' dependent score in the j th condition, subject to the constraint that all α_j sum to zero across all treatment conditions of Factor A .
$\pi_{i(j)}$	is the population effect associated with each block i (i.e., subjects in the above diagram) nested within the j th condition of Factor A . Its underlying population distribution is assumed to be normal with equal variance in all blocks.

β_k	is the effect for the k th treatment condition of Factor B ; algebraically, it equals the deviation of the population mean ($\mu_{\cdot k}$) from the grand mean (μ). It is a constant for all subjects' dependent scores in the k th condition, subject to the constraint that all α_j sum to zero across all treatment conditions of Factor B .
$(\alpha\beta)_{jk}$	is the effect due to the combination of level j of Factor A and level k of Factor B . It is subject to the constraint that all $(\alpha\beta)_{jk}$ sum to zero across all treatment conditions of Factor A and Factor B .
$(\beta\pi)_{ki(j)}$	is the effect due to the combination of level k of Factor B and block i , both nested within the j th condition of Factor A . It is assumed that the underlying distribution of $(\beta\pi)_{ki(j)}$ is normal with equal variance across all blocks and all levels of Factor B . Furthermore, $(\beta\pi)_{ki(j)}$ is assumed to be statistically independent of the effect $\pi_{i(j)}$.
ε_{ijk}	is the random error effect associated with the i th block nested within the j th level of Factor A and crossed with the k th condition of Factor B . It is a random variable that is normally distributed in the underlying population and is independent of the effect $\pi_{i(j)}$.

To test statistical hypotheses concerning the effects of Factor A , Factor B , or A and B interaction, two error terms are needed. The first error term is used in testing the between-subjects factor (i.e., A) and is computed from the average of (two, in the example above) mean squares within A levels by ignoring Factor B . The second error term is used in testing within-subjects factor (i.e., B) and its interaction with the between-subjects factor (i.e., A by B). It is computed from the average of (four, in this example) mean squares within the A by B combinations. Technically, it is the pooled interaction of Factor B and the blocks (i.e., subjects, in this example) that are subsumed under each level of Factor A . All statistical tests are performed by the F statistic under the assumption of homogeneity of variance. The F tests of the within-subjects factor and its interaction with the between-subjects factor require an additional assumption, namely, the sphericity assumption for the variance and covariance matrix of the within-subjects scores. The sphericity assumption is the necessary and sufficient condition for the validity of the F test (Huynh & Feldt, 1970).

—Chao-Ying Joanne Peng

See also REPEATED-MEASURES DESIGN, NESTED DESIGN, BLOCK DESIGN

REFERENCES

- Hays, W. L. (1994). *Statistics*. Fort Worth, TX: Holt, Rinehart and Winston.
- Huck, S. (2003). *Reading statistics and research* (4th ed.). New York: Pearson Education Inc.
- Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association*, 65, 1582–1589.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.
- Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.
- Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

S-PLUS

S-Plus is a general-purpose statistical software package that relies on object-oriented programming. It is a powerful and flexible computer program that can be easily tailored for individual purposes. For further information, see the website of the software: <http://www.insightful.com/products/default.asp>.

—Tim Futing Liao

SPREAD

Spread of data refers to how the data are distributed. Examples of the measures of spread include RANGE, VARIANCE, and STANDARD DEVIATION.

—Tim Futing Liao

See also DISPERSION

SPSS

SPSS (Statistical Package for the Social Sciences) is a general-purpose software package that is popular among social scientists and satisfies many of their

computing needs. For further information, see the website of the software: <http://www.spss.com>.

—Tim Futing Liao

SPURIOUS RELATIONSHIP

A spurious relationship implies that although two or more variables are correlated, these variables are not causally related. CORRELATION analysis merely establishes covariation, the extent to which two or more phenomena vary together, whereas CAUSATION concludes that one phenomenon caused another. Although social scientists are interested in correlation, most would also like to infer causality. For example, although we know that smoking is related to lung cancer, we want to infer the possibility that smoking causes lung cancer. But when seeking inferences of causality, social scientists must beware the possibility of spurious relationships, which falsely imply causation.

Suppose we are estimating a coefficient of correlation r_{xy} , where x is the marital status of an individual and y is the amount of wine consumed annually by that individual. We find a high positive correlation between marital status and wine consumption. If we infer a causal relationship, we may posit that consuming wine causes people to get married. We may also posit that marriage drives individuals to drink wine. Logically, neither of these interpretations makes sense. It is more likely that the relationship between wine consumption and marriage is spurious, and a third variable is causing positive correlation in both variables.

Spurious correlation means that the relationship between two variables disappears when a third variable is introduced. Theory or logic should specify a third variable that may account for the observed correlation between the variables of interest. In the example above, we suspect that variable z , age, may have a joint effect on both marital status and wine consumption. We find a high positive correlation, r_{xz} , between marital status and age, and a high positive correlation, r_{yz} , between wine consumption and age. These strong correlations offer initial evidence that variation in age accounts for the relationship between marital status and wine consumption. To test for spuriousness, we calculate $r_{xy \cdot z}$, the PARTIAL CORRELATION COEFFICIENT between marital status and wine consumption, which CONTROLS for age, where $r_{xy \cdot z} = r_{xz} * r_{yz}$. We find that $r_{xy \cdot z}$

does not significantly differ from zero, meaning that the relationship between marital status and wine consumption disappears when controlling for age. We thus speculate that the correlation between marital status and wine consumption results from the joint effect of variation in age. In other words, the relationship between marital status and wine consumption could be spurious.

Because the value for $r_{xy \cdot z}$ does not significantly differ from zero, we might also conclude that z is a MEDIATING VARIABLE between x and y . In the above example, this would imply that marriage leads to aging, and becoming older leads to more wine consumption. For our example, this is a nonsensical conclusion, but in other cases, the distinction may not be clear. It is up to the researcher to make logical assumptions about the direction of causation and to use those assumptions to conclude if x and y are spuriously related, or if z is a mediating variable. Statistical evidence alone cannot distinguish between these two possibilities.

—Megan L. Shannon

REFERENCES

- Blalock, H. M., Jr. (1979). *Social statistics* (2nd ed.). New York: McGraw-Hill.
- Simon, H. A. (1954). Spurious correlation: A causal interpretation. *Journal of the American Statistical Association*, 49, 467–479.

STABILITY COEFFICIENT

Stability coefficient is a term that, in the social sciences, means the correlation of measurement results from Time 1 with measurement results from Time 2, where the subjects being measured and the measuring instrument remain precisely the same. If the coefficient is high (and no instrument reactivity is present), it generally signals both measurement RELIABILITY and response continuity. If it is low, a correct interpretation is harder to reach.

A low stability coefficient with measures of personal history or behavior plainly derives from unreliability. If respondents first report and then later deny features of personal background or past conduct, one does not interpret the change as “discontinuity.” However, with attitudes and opinions, the low coefficient is of less certain meaning. It is not uncommon for such a

coefficient to be treated as an indicator of unreliability when the test-retest interval is short (Carmines & Zeller, 1979; Litwin, 1995) but as an indicator of attitudinal discontinuity when that interval is long (Converse & Markus, 1979). Neither of these two interpretations is self-evidently correct, and disentangling the reliability and continuity components of a stability coefficient can be complicated and important (Achen, 1975). Indeed, an accurate estimate of continuity cannot be determined without knowing, and correcting for, the level of measurement unreliability.

A TEST-RETEST time interval of days or weeks obviously strengthens the assumption that low stability is caused by measurement problems rather than attitude change. Even here, however, other factors may intervene. For example, in turbulent times, there may be considerable short-term shuffling of true opinions. And whether the times be turbulent or peaceful, accurately stated opinions on unfamiliar or ill-considered topics also may show substantial short-run instability. Attitudinal ephemera do not represent true measurement error, although if one embraces a definition of “attitude” that emphasizes its enduring nature, such ephemera may be regarded as “nonattitudes.”

The use of CORRELATION coefficients, whether product-moment or rank-order, in estimating continuity may sometimes be problematic. With either standardized or ranked data, there can be systematic change between Time 1 and Time 2 that is not reflected in the correlations. Response distributions may shift upward or downward or may tighten or spread, and stability coefficients nonetheless remain high. For that reason, it is important to look further into the correlation, inspecting scatter diagrams with interval data or contingency tables with ranked data, for signs of such distributional changes.

—Douglas Madsen

REFERENCES

- Achen, C. H. (1975). Mass political attitudes and the survey response. *American Political Science Review*, 69, 1218–1231.
- Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.
- Converse, P. E., & Markus, G. B. (1979). Plus ça change . . . : The new CPS Election Study Panel. *American Political Science Review*, 73, 32–49.
- Litwin, M. S. (1995). *How to measure survey reliability and validity*. Thousand Oaks, CA: Sage.

STABLE POPULATION MODEL

What are the long-term implications of the maintenance of current demographic patterns over a long period? Alfred Lotka's extension of biological models of population renewal to the domain of age-specific rates provides many of these answers (Lotka, 1939/1998). A stable population, having constant birth rates and age distribution, emerges when the following three characteristics, which closely approximate the long-term conditions of many historic populations, persist over the long term:

1. Age-specific fertility rates are constant.
2. Age-specific death rates (i.e., the LIFE TABLE) are constant.
3. Age-specific net migration rates are zero.

The simple demographic properties underlying stable populations form the basis of population projections and the analysis of demographic components (age distribution, mortality, and fertility). Of particular importance is the property of *ergodicity*, which states that the long-term age distributions, birth rates, death rates, and growth rates of stable populations are determined by their current fertility and mortality schedules irrespective of current age distributions.

The long-term size, population distribution, and vital rates of a stable equivalent population can be predicted from current vital rates (see DEMOGRAPHIC METHODS for more). The proportion of the population at each age (${}_nK_x^{\text{stable}}$) can be represented in terms of life table survival functions (${}_nL_x$), births (B), and population growth (r)

$${}_nK_x^{\text{stable}} = B {}_nL_x e^{-r \cdot x}.$$

The total number of births, B , varies according to the initial population age distribution, yet long-term crude birth rates and yearly growth in the number of births are functions only of fertility and women's mortality schedules. The *intrinsic growth rate*, r , or the stable growth rate implied by current vital rates, bears a direct and easily estimated relationship to the Net Reproduction Rate (NRR), the average number of girls produced by each woman based on current birth and death rates:

$$\text{NRR} = e^{rT},$$

where T is the mean length of a generation.

Figure 1 depicts the stabilization process through a simple hypothetical population projection (see LESLIE'S MATRIX). Italy's vital rates from 1995, characterized by low fertility and low mortality, are used to project forward the 1995 age distributions of Italy (a relatively aged society) and Nigeria (a young society). Their age distributions differ dramatically over the medium term (2030–2060), as large cohorts of very young Nigerians age into their childbearing years, producing relatively large cohorts of children in spite of low fertility rates. Over the long term (2100–2200), the age distributions grow statistically identical, as the ripple effects of earlier cohorts in both populations diminish.

Although their age distributions are identical by 2200, Nigeria's overall population growth would have been much greater. The stable population model can be applied to the measurement of *population momentum*, the population growth and temporary demographic changes emerging when fertility decline occurs in a population with a young age distribution (Keyfitz, 1985). Although Nigeria's population would have grown substantially more than Italy's as its largest ever cohorts aged into childbearing, all of this growth would have occurred at ages above the mean age of childbearing; new cohorts could only be smaller than old cohorts given such low fertility rates.

Simplified demographic accounting procedures stemming from the stable population model have led to a greater empirical understanding of the complex medium- and long-term implications of demographic change. In particular, family planning strategies have been influenced by the substantial impact of *delayed* childbearing on future population size (e.g., "later, longer, fewer") through associated increases in the mean age of childbearing. Government pension schemes have been influenced by the greater relative impact of fertility decline, vis-à-vis mortality decline, on population aging and old-age dependency (the relative number of elders compared to young adults), and the cyclical economic effects of rapid population aging as large cohorts age from childhood dependency, to adult productivity, to old-age dependency.

Further methodological innovations have stemmed from relaxing key assumptions of the stable population model. The *variable- r method* generalizes the properties of stable populations to nonstable populations, facilitating a wide range of demographic projection and correction techniques (Preston & Coale, 1982). Generalization of the stable model to populations

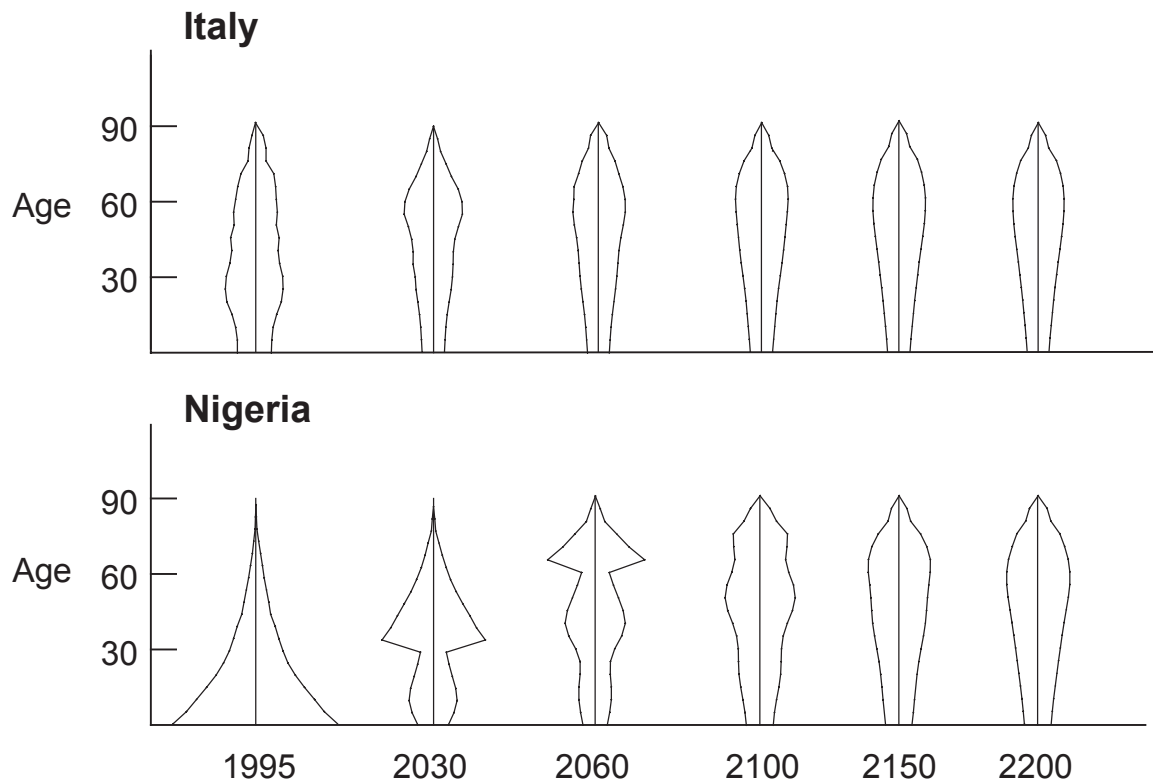


Figure 1 Relative Age Distributions for Italy and Nigeria, Both Projected Using Vital Rates for Italy, 1995
 SOURCE: Adapted from Preston, Heuveline, and Guillot (2001) and United Nations (1995, 1996).

with constant, nonzero, age-specific migration rates facilitated the development of multistate and multi-regional stable population models (Rogers, 1995).

For further discussion and examples, see Preston et al. (2001).

—Randall Kuhn

REFERENCES

- Keyfitz, N. (1985). *Applied mathematical demography* (2nd ed.). New York: Wiley.
- Lotka, A. J. (1998). *Analytical theory of biological populations: Part II: Demographic analysis with particular application to human populations* (D. P. Smith & H. Rossert, Trans.). New York: Plenum. (Original work published 1939)
- Preston, S. H., & Coale, A. J. (1982). Age structure, growth, attrition and accession: A new synthesis. *Population Index*, 48(2), 217–259.
- Preston, S. H., Heuveline, P., & Guillot, M. (2001). *Demography: Measuring and modeling population processes*. Oxford, UK: Blackwell.
- Rogers, A. (1995). *Multiregional demography*. Chichester, UK: Wiley.

United Nations. (1995). *World population prospects: The 1994 revision*. New York: Author.

United Nations. (1996). *Demographic yearbook 1994*. New York: Author.

STANDARD DEVIATION

One of the primary properties of a set of numbers or scores is its variability or DISPERSION—how much the scores differ from one another. Probably the most commonly used measure of the variability of a set of scores is the standard deviation. It is popular both because it takes into account the exact numerical value of every score and because it is defined in the units of the scores themselves. That is, if heights are measured in inches, the standard deviation is also expressed in inches.

Most often, the standard deviation is considered as a supplement to measures of average, such as the arithmetic MEAN. For example, many cities have

similar mean temperatures for the year, but they differ widely in variability of temperatures across seasons. In describing and comparing the weather in these cities, it would be informative to classify them in terms of variations in temperature across the year as well as overall yearly average temperature.

Similarly, in research designed to generate appropriate inferences about human behavior, standard deviations can be as important as means. Consider, for example, a program designed to increase students' test performance. To be sure, you'll be looking for an increase in average scores following the program. However, this increase could come in several forms. If the program helped each student to approximately the same degree, the mean scores would increase from before to after the program, but the standard deviations would be about the same. If, however, the program helped only the best students while leaving the rest unchanged, you would still get an increase in mean but also an increase in standard deviation (i.e., a greater separation of the best and worst students). For some educators, this latter result would not be considered as desirable as would be evident from inspection of means alone. The increased standard deviation would serve as an important clue as to how the program affected scores.

The standard deviation of a complete set or population of scores, σ , is defined as the square root of the sum of the squared differences between each score X and the mean of the scores, μ , divided by the number of scores N ; that is,

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}.$$

Squared differences are used because otherwise, the positive and negative differences would cancel each other. Without the square root, the measure σ^2 is called the VARIANCE, but it is expressed in units squared (e.g., square inches) rather than the units themselves. In computing the standard deviation of a sample of scores, s , as an estimate of the population standard deviation, a slight correction is made in the formula:

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}},$$

where $\bar{X}$ is the mean of the sample of scores and N is replaced by $N - 1$. This corrects for an overall bias to underestimate the population

standard deviation based on data from a sample. A "computational" form of the same equation is the following:

$$s = \frac{\sum X^2 - [(\sum X)^2 / N]}{N - 1}.$$

To illustrate, consider the following sample of 10 scores:

8, 26, 15, 9, 13, 20, 21, 10, 15, 12.

For these scores, $\sum X = 149$; $(\sum X)^2 = 22,201$; $\sum X^2 = 2,525$; and $s = 5.82$. A good summary of this sample of 10 scores is that it has a mean of 14.90 and a standard deviation of 5.82. Without the standard deviation, this description would be incomplete.

—Irwin P. Levin

STANDARD ERROR

The standard error of a STATISTIC is the STANDARD DEVIATION of the SAMPLING DISTRIBUTION of the statistic, and it is used to measure the RANDOM VARIATION of that statistic around the PARAMETER that it estimates. For example, the standard error of a MEAN is the standard deviation of a DISTRIBUTION of means for an infinite number of samples of a given size n . The size of the standard error is dependent, in part, upon the sample size, and in accordance with the Law of Large Numbers, as sample size increases, the standard error decreases. In addition, if a statistic is an UNBIASED ESTIMATOR of its parameter, the larger the sample size, the more closely the statistics will cluster around the parameter.

The true standard error of a statistic is generally unknown, and the term *standard error* commonly refers to an estimate. Although standard errors can be estimated from a FREQUENCY DISTRIBUTION by repeated sampling, usually they are estimated from a single sample and represent the expected standard deviation of a statistic as if a large number of such samples had been obtained. Estimates of standard errors from a single sample are possible because the properties of the sampling distributions for certain statistics have already been established. Methods for estimating a standard error from a single sample are demonstrated

below for two statistics: a sample mean, $\bar{X}$, and a REGRESSION COEFFICIENT, $\hat{\beta}_1$.

THE STANDARD ERROR OF A SAMPLE MEAN, $\bar{X}$

The CENTRAL LIMIT THEOREM establishes that the sampling distribution of $\bar{X}$ is normally distributed if the sample was drawn from a population that is itself normally distributed, even if the sample size is small. Furthermore, if the sample size is large, the sampling distribution of $\bar{X}$ is normally distributed, even if the population is not normally distributed. Thus, given the central limit theorem, virtually every $\bar{X}$ is assumed to have a sampling distribution that is normally distributed, and because $\bar{X}$ is an unbiased estimator of the population mean μ , the mean of the sampling distribution is equal to μ , and the standard error, $\sigma_{\bar{X}}$, is equal to the population standard deviation of X , σ , divided by the square root of the sample size:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

When the population standard deviation is unknown, as is generally the case, the standard error is estimated using the sample standard deviation, s , or

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}.$$

In other words, $\sigma_{\bar{X}}$ is the standard error of $\bar{X}$, and $s_{\bar{X}}$ is the estimated standard error of $\bar{X}$.

THE STANDARD ERROR OF A REGRESSION COEFFICIENT, $\hat{\beta}_1$

If the DEPENDENT VARIABLE (Y) in a REGRESSION model is drawn from a sample of normally distributed observations that are independent and identically distributed, ORDINARY LEAST SQUARES estimates of the parameters have normally distributed sampling distributions. In a simple regression model with only one INDEPENDENT VARIABLE, $Y = \beta_0 + \beta_1 X + \varepsilon$, the estimate of β_1 is denoted $\hat{\beta}_1$. The sampling distribution of $\hat{\beta}_1$ is normally distributed with mean equal to the parameter β_1 and estimated variance $s_{\hat{\beta}_1}^2$ equal to the estimated variance of the RANDOM ERROR term s^2 divided by the sum of the squared deviations of X . Thus, the estimated standard error of $\hat{\beta}_1$ is the square root of $s_{\hat{\beta}_1}^2$, or

$$s_{\hat{\beta}_1} = \sqrt{\frac{s^2}{\sum (X - \bar{X})^2}}.$$

The numerator indicates that the standard error of $\hat{\beta}_1$ is smaller when the error variance is smaller, and the denominator indicates that more variance in X results in a smaller standard error of $\hat{\beta}_1$, or a more precise estimator.

—Jani S. Little

REFERENCES

- Devore, J. L. (1991). *Probability and statistics for engineering and the sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
- Hamilton, L. C. (1990). *Modern data analysis: A first course in applied statistics*. Pacific Grove, CA: Brooks/Cole.
- Kmenta, J. (1986). *Elements of econometrics* (2nd ed.). New York: Macmillan.
- Sokal, R. R., & Rohlf, F. J. (1981). *Biometry* (2nd ed.). San Francisco: W. H. Freeman.

STANDARD ERROR OF THE ESTIMATE

The standard error of the estimate (SEE) is, roughly speaking, the average “mistake” made in predicting values of the DEPENDENT VARIABLE (Y) from the estimated regression line. The SEE is assumed to be constant over all values of X . Thus, the average error in predicting Y when $X = x_i$ and $X = x_j$, $x_i \neq x_j$, will be the same. This is implied by the assumption of constant error variance [$\text{Var}(u_i|X_i) = \sigma^2$] in the classical linear regression model (CLRM).

There is an intuitive similarity between the SEE and the STANDARD DEVIATION of a RANDOM VARIABLE. The standard deviation of a random variable is nothing more than the square root of the average squared deviations from its MEAN. Similarly, the SEE is the square root of the average squared deviations from the regression line. Within a regression framework, these deviations are represented by the DISTURBANCE TERMS [$\hat{u}_i = (y_i - \hat{y}_i)$]. Thus, the SEE can be understood as the standard deviation of the SAMPLING DISTRIBUTION of disturbance terms, which, by assumption, is centered on zero. The SEE is also known as the root MEAN SQUARE ERROR of the regression.

For a regression in which k parameters are estimated, where k is taken to be the number of independent variables in the regression model plus the constant, the SEE is given by the following:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - k}}.$$

The SEE is a measure of the GOODNESS OF FIT of the estimated REGRESSION line to the data. The smaller the SEE, the better the regression line fits the data. If the regression line fits the data perfectly, then each observation in the data set will fall exactly on the regression line and the SEE will be zero. Some researchers consider the SEE to be the preferred measure of fit of a regression model, and the statistic has many advantages. First, it is expressed in units of the dependent variable allowing for meaningful comparisons across regressions with the same dependent variable. Also, it is not dependent on the VARIANCE of the INDEPENDENT VARIABLES in the model as is another commonly used measure of fit, R^2 .

To see how the SEE is calculated, consider an example. An instructor is interested in predicting student performance in an introductory statistics class. The instructor believes that students' final exam grades are a LINEAR additive function of the following independent variables: (a) GPA, (b) SAT, (c) Gender, and (d) Year in School. Using a RANDOM SAMPLE of 15 students, the instructor generates the following PREDICTION EQUATION:

$$\hat{Y}_i = 9.14 + 1.78GPA + 0.11SAT \\ - 2.67GENDER + 2.10YEAR.$$

These estimates are then used to generate predicted values on Y for each of the students sampled (see Table 1).

How well does the estimated regression line fit the observed data? To answer this question, the instructor calculates the SEE. Using the data in Table 1 and the formula given above the SEE,

$$\hat{\sigma} = \sqrt{\frac{144.69}{15 - 5}} = \sqrt{14.47} = 3.80.$$

On average, the model will make a prediction error of 3.80 points. Recalling that a small standard error is associated with a better fit (and noting that the standard deviation of Y is 9.68), we would conclude that the estimated regression line fits the data in this example well.

—Charles H. Ellis

REFERENCES

Achen, C. H. (1982). *Interpreting and using regression*. Beverly Hills, CA: Sage.

Table 2 Final Scores, Predicted Scores, and Prediction errors

	Final Grade (Y_i)	$\hat{Y}$	$u_i = (y_i - \hat{y}_i)$	u_i^2
1	82	85.02	-3.02	9.11
2	90	85.43	4.57	20.90
3	78	81.39	-3.39	11.46
4	68	65.50	2.50	6.27
5	76	78.51	-2.51	6.630
6	69	74.93	-5.93	35.14
7	91	89.62	1.38	1.90
8	82	83.00	-1.00	1.00
9	91	88.69	2.31	5.33
10	66	66.69	-0.69	0.48
11	77	75.53	1.47	2.16
12	58	58.23	-0.23	0.05
13	75	69.64	5.36	28.78
14	85	82.64	2.36	5.59
15	74	77.20	-3.20	10.23
			$\sum_{i=1}^n u_i = 0$	$\sum_{i=1}^n u_i^2 = 144.69$

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.

STANDARD SCORES

Standard scores enable us to more readily understand the meaning of a particular score. For example, knowing that a student answered 20 items correctly on a 30-item math test gives us only a very rough view of that student's performance. It is hard to know how good or bad this score is without knowing how other students at the same grade level scored. We would know more if we were told that the mean score for all other students at the same grade level taking the exam under the same circumstances was 17. We would at least know that our student was "somewhat above average." If we were also told that the scores had a standard deviation of 3, we would have a better idea of "how much above average," namely, one standard deviation above the mean. Suppose we learned that the same student answered 35 items correctly on a 50-item test of verbal ability where the mean was 40 and the standard deviation was 5. We would then know that

this same student scored one standard deviation below the mean on the verbal test.

As illustrated in the above example, standardizing sets of test scores by adjusting each score to the mean and STANDARD DEVIATION of the set serves the following two important functions: (a) It allows us to relate one student's test score to the complete set of scores for all comparable students, and (b) it allows us to compare the score on one test to the score on another test for the same student. The latter can be particularly useful in identifying a student's relative strengths and weaknesses.

There are a variety of ways of forming standard scores. For example, some personality measures are adjusted such that the mean will be 50 and the standard deviation 10, and some intelligence tests are scored such that the mean will be 100 and the standard deviation 15. The most common form of standard score used by social scientists, however, is the Z-SCORE, which transforms each score X by subtracting the mean of the scores, $\bar{X}$, and dividing the difference by the standard deviation, s : $z = (X - \bar{X})/s$.

Each TRANSFORMED score then directly represents how many standard deviations above or below the mean that score is. $z = +1$ means the score is one standard deviation above the mean, $z = -1$ means one standard deviation below the mean, $z = 0$ means that that score falls at the exact arithmetic mean, and so forth. When the form of the original distribution of scores is known, these standard scores can be further translated into PERCENTILES. For example, in a NORMAL DISTRIBUTION, only 16% of the scores are more than one standard deviation above the mean, so a z -score of $+1$ falls at the 84th percentile. (Most statistics books include a table for translating standard scores in a normal distribution into percentiles.)

Another useful property of z -scores and other standard scores formed by adjusting for the mean and standard deviation is that they are unit-free. For example, a person's standard score for height does not depend on whether height was measured in inches or in centimeters. Likewise, CORRELATION is calculated from standard scores so that, for example, one could compare the correlation between years of education and income across cultures without concern for different monetary units in different countries. The standardization implied by the term *standard score* has many practical advantages.

—Irwin P. Levin

STANDARDIZED REGRESSION COEFFICIENTS

In ORDINARY LEAST SQUARES (OLS) MULTIPLE REGRESSION analysis, an *UNSTANDARDIZED* regression coefficient, b , describes the relationship between a DEPENDENT and an INDEPENDENT VARIABLE in terms of the original *units* of measurement (dollars, kilograms, scale scores, years) of those variables. A one-unit change in the independent variable is expected to produce b units of change in the dependent variable. For example, a 1-year increase in the number of years of school completed may result in a $b = \$1,024$ increase in annual income. A *standardized* coefficient, b^* , expresses the same relationship in STANDARD DEVIATION units: A one-standard-deviation increase in the number of years of school completed may result in a $b^* = .275$ standard deviation increase in annual income. In OLS regression, the standardized coefficient b^* can be calculated in either of two ways. One is to TRANSFORM all variables to Z-SCORES or STANDARDIZED VARIABLES (from each variable, subtract its mean, then divide the result by its standard deviation) prior to the REGRESSION analysis. The other is to multiply the unstandardized coefficient for each independent variable by the standard deviation of that independent variable, then divide by the standard deviation of the dependent variable: $b^* = (b)(s_x)/(s_y)$, where y is the dependent variable, x is the independent variable, s_y is the standard deviation of y , s_x is the standard deviation of x , and b and b^* are the unstandardized and standardized coefficients, respectively, for the relationship between x and y .

Standardized coefficients are used for two main purposes. First, when independent variables are measured in different units (for example, years of schooling for one variable, IQ score for another, etc.), it is difficult, based only on the unstandardized coefficients, to discern which of two or more independent variables has more impact on the dependent variable. The same variable measured in different units (centimeters, meters, or kilometers) will have a larger or smaller unstandardized regression coefficient. Using standardized coefficients converts all of the coefficients to a common unit of measurement (standard deviations), and the standardized coefficient is the same regardless of the units in which the variable was originally measured. A larger standardized coefficient for one independent variable as opposed to a second

independent variable allows us to conclude that the first independent variable has a stronger influence than the second independent variable, even when the two independent variables were originally measured in very different units. Second, standardized coefficients are used in PATH ANALYSIS to calculate indirect and total effects, as well as direct effects, of an independent variable on a dependent variable. In a *fully RECURSIVE* path model, it is possible to “decompose” the correlation between an independent variable and a dependent variable into the direct influence and the indirect influence (the independent variable influences other variables, which in turn influence the dependent variable) of the independent variable. In the special case of a single independent variable, the standardized regression coefficient is equal to the correlation between the two variables.

Moving beyond OLS multiple regression, Goodman (1973a, 1973b) constructed what he called a “standardized” coefficient for LOGIT MODELS by dividing the unstandardized coefficient by its standard error, but this is really just a variant of the Wald statistic sometimes used to test the STATISTICAL SIGNIFICANCE of LOGISTIC REGRESSION coefficients. Goodman also attempted to construct a parallel to path analysis using the logit model, but the path analysis analogue he proposed did not permit the calculation of indirect effects or the decomposition of CORRELATIONS between dependent and independent variables (Fienberg, 1980). More recently, Menard (1995) reviewed other attempts to construct standardized logistic regression coefficients and suggested a procedure for a *fully standardized logistic regression coefficient*, standardized on both the dependent and the independent variables, as opposed to partially standardized logistic regression coefficients, which are standardized only on the independent variable. Menard (1995) and work in progress by Menard (in press) suggests that this fully standardized logistic regression coefficient can be interpreted in much the same way as the standardized OLS regression coefficient, and that it can be used in path analysis with logistic regression, including calculation of indirect effects and decomposition of correlations between dependent and independent variables.

—Scott Menard

REFERENCES

- Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge: MIT Press.
- Goodman, L. A. (1973a). The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika*, 60, 179–192.
- Goodman, L. A. (1973b). Causal analysis of data from panel studies and other kinds of surveys. *American Journal of Sociology*, 78, 1135–1191.
- Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.
- Menard, S. (in press). *Logistic regression*. Thousand Oaks, CA: Sage.

STANDARDIZED TEST

Standardized tests are used to measure cognitive skills, abilities, aptitudes, and achievements in a variety of applications in psychology, education, and other social sciences. Standardized MEASURES also have been developed to assess personality, attitudes, and other noncognitive characteristics. However, the focus of the present entry is on the measurement of cognitive characteristics, which are often referred to as tests.

A standardized test is a measure of a well-articulated CONSTRUCT, has clearly stated purposes, is developed using highly structured procedures, is administered under carefully controlled conditions, has a systematic basis for score interpretation, and is shown to possess various technical properties.

CONSTRUCT DEFINITION

Standardized tests are measures of constructs. The AERA/APA/NCME (1999) *Test Standards* define a construct “broadly as the concept or characteristic that a test is designed to measure” (p. 5). In aptitude testing, a test might be constructed to measure the construct spatial ability. In achievement testing, a test might be developed to measure the construct mathematics achievement. A major component of standardized tests is the construct definition, which is a clear statement of the construct that is to be measured by the test. A construct definition provides a foundation for developing a test to measure that construct and for validating the test.

TEST DEVELOPMENT

Systematic procedures are used to develop standardized tests. Test development procedures help provide an OPERATIONAL definition of how a test assesses

the construct. For standardized aptitude tests, the different item types included on the test are specified, as are the rules for developing items. For standardized achievement tests, detailed tables of specifications are developed that indicate the item types, content areas, and cognitive skill levels of the items to be included. Such test specifications provide test developers with a blueprint for test construction.

Standardized tests also have detailed statistical specifications. Pretest procedures are used to understand how test items function for examinees before the test items are included on the operational versions of the tests. Statistical indexes of item difficulty and discrimination are often calculated based on pretesting. Statistical specifications indicate how many items of various levels of difficulty and discrimination are included on the test.

The use of detailed content and statistical specifications is a requisite condition for the development of alternate forms of test, which are different sets of test items that can be expected to produce scores with very similar meanings and statistical properties. Alternate forms enhance test security and allow tests to be administered to an examinee more than one time.

TEST ADMINISTRATION

Standardized tests have strictly controlled administration conditions: Examinees are given explicit instructions, test booklets or other testing materials are provided in the same manner to all examinees, time allowed for testing is carefully controlled, the same answer documents or procedures are used for all examinees, and the use of other tools such as calculators and scratch paper is controlled and specified.

EXAMINEE RESPONSES

Many standardized tests use a multiple-choice format that requires examinees to indicate which one from a set of possible answers is the correct one. Other objectively scored item formats also exist, such as true-false. On some tests, examinees might be asked to provide short answers to questions or extended answers, such as with an essay test, and on other tests, examinee behavior is observed.

SCORING PROCEDURES

Well-defined scoring procedures are used with standardized tests. For objectively scored tests, a raw

score often is defined as the number of items that the examinee correctly answered. Other, more complex raw scores, such as partial credit scores, are sometimes used. For tests in which examinees provide short or extended answers, human judges are used to score the responses. In such cases, detailed scoring rubrics are used that provide explicit guidelines to the judges on how to score the responses.

SCORE SCALES

Raw scores on standardized tests are TRANSFORMED to SCALE scores to facilitate score interpretation. Score scales can be normatively based. For example, norming studies are conducted for intelligence tests in which the test is administered to a group of examinees that is intended to represent a particular population, such as, for example, 14-year-olds in the United States. The mean raw score from the norming study might be set equal to a scale score of 100. Other normative points are used to set other points on the score scale.

Proficiency-level scoring sometimes is used with educational achievement tests. In proficiency-level scoring, a standard setting study is conducted in which, through a systematic procedure, judges define the raw score an examinee needs to be at or above for a particular proficiency level. Often, a series of proficiency levels is defined in this process. One set of frequently used proficiency levels is *basic*, *proficient*, and *advanced*.

Norms and proficiency levels provide test users a foundation for interpreting scores. As such, they are a major component of standardized tests.

TEST SECURITY

It is important that the scores on the test represent the construct that is being measured. However, if the test questions become known to future examinees, performance might not reflect the examinees' proficiency on that construct. For this reason, many standardized testing programs have elaborate test security procedures. The procedures include using alternate forms, keeping tight control over testing materials, and sometimes running statistical analyses after test administrations to identify irregularities in security.

RELIABILITY

Standardized tests are developed to produce scores that have sufficient RELIABILITY for making important

decisions. Decisions about testing time, numbers of test questions, and types of questions to include on the test are made to ensure that the scores have sufficient reliability for their intended purposes.

VALIDITY

Test VALIDITY is the degree to which evidence supports interpretations of test scores. A test validation process involves collecting evidence that provides a sound scientific basis for score interpretations. Validation begins with statements of the construct to be measured, the purposes of the test, and the interpretations to be made of the scores. Various types of evidence are collected to support the validity of the test. Sources of validity evidence can be based on test content, examinee response processes, and a study of the internal psychometric structure of the test, as well as by relating the test scores to other variables. Evidence that the test is not influenced by variables extraneous to the construct being measured is an important component of a validation process. A sound validity process integrates the different sources of evidence to support the intended interpretation of test scores for specific uses.

SPECIAL ISSUES

In recent years, the administration of some standardized tests has moved to computers, creating new challenges for the development of standardized tests. PSYCHOMETRIC test theories, such as ITEM RESPONSE THEORY and GENERALIZABILITY THEORY, have seen major advances in recent years, leading to better understanding of the technical properties of standardized tests and making possible new procedures for administering and scoring standardized tests.

—Michael J. Kolen

REFERENCES

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anastasi, A., & Urbina, S. (1996). *Psychological testing* (7th ed.). New York: Prentice Hall.
- Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). New York: Macmillan.
- Plake, B. S., & Impara, J. C. (Eds.). (2001). *The fourteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

STANDARDIZED VARIABLE. See Z-SCORE; STANDARD SCORES

STANDPOINT EPISTEMOLOGY

Standpoint epistemology developed from feminist criticisms regarding women's absence from, or marginalized position in, social science. Although social science was supposedly OBJECTIVE and value free, feminists argued that it was conducted largely from male perspectives and male interests. In order to make the meaning of women's lives more visible, it was necessary to analyze it from their point of view. The significance of standpoint epistemology lies in its challenge to descriptions and classifications of social life that are based on universalistic male assumptions. Key is the idea that women and men lead lives with differing contours, creating different kinds of knowledge. Women's experience of oppression produces particular and privileged understandings. This access to different knowledge is able to reveal the existence of forms of human relationships that may not be visible from the position of the ruling male gender. Therefore, standpoint epistemology offers the possibility of new and more reliable insights into gendered power and relationships. It has significance for the practice of FEMINIST RESEARCH, for example, through emphasizing the importance of listening to women's voices. It also has possibilities for extension into understanding the lives of other (e.g., minority, ethnic, and disabled) groups.

Variations exist in how a feminist standpoint might be conceived and its implications for knowledge, with some using the expression *theory*, rather than *epistemology*, because of the latter's association with certainty and neutrality. It is also suggested that standpoint is a practical achievement, rather than an ascribed status; subordination alone does not lead to an awareness of inequality and oppression. One major difficulty with a standpoint approach is that it risks replacing a generalized male perspective with that of a generalized woman. Recently, feminists have questioned the idea that the nature of womanhood is fundamental or intrinsic, and that women's oppression is universal, homogeneous, and shared. They have moved to an emphasis on difference and diversity among women,

particularly focusing on the structural positions derived from ethnicity, sexuality, disability, class, and age. This has led to a proliferation of differing standpoints, for example, those of some Black and lesbian feminists. Yet such an approach also raises problems about the existence of a multiplicity of incommensurable standpoints and of RELATIVISM. It assumes that one structural position alone, rather than multiple positions—which may invoke conflict or tensions—provides particular knowledge. There is also the risk that orthodoxies develop around particular standpoints and ways of being and seeing, so that some women might be judged as deviant from an assumed norm.

The usefulness and implications of standpoint epistemology have been extensively debated, and disagreements abound. However, there is some consensus that it highlights the relationship between women's experiences and feminist knowledge and between knowledge and power. There is no such thing as a disinterested knowledge, and adopting a standpoint draws attention to the subject at the center of knowledge production.

—Mary Maynard

REFERENCES

- Collins, P. H. (1997). Comment on Heckman's "Truth and method: Feminist standpoint theory revisited": Where's the power? *Signs*, 22(21), 375–381.
- Hartsock, N. C. M. (1998). *The feminist standpoint revisited and other essays*. Boulder, CO: Westview.
- Smith, D. (1997). Comment on Heckman's "Truth and method: Feminist standpoint theory revisited." *Signs*, 22(21), 392–397.

STATA

Stata is a general-purpose statistical software package that has gained popularity in recent years in the social sciences, at least in part because of its flexibility in implementing many user-written, nonetheless well-maintained, macros. For further information, see the Web site of the software: <http://www.stata.com>.

—Tim Futing Liao

STATISTIC

A *statistic* is simply a numerical summary of the SAMPLE DATA. For example, data analysts often

use familiar sample statistics such as sample MEANS, sample VARIANCES and STANDARD DEVIATION, sample PERCENTILES and MEDIANS, and so on to summarize the samples of QUANTITATIVE data. For samples of CATEGORICAL data, sample proportions, percentages, and ratios are typical sample statistics that can be utilized to summarize the data. In short, a statistic is simply a summary quantity that is calculated from a sample of data.

The counterpart of a statistic for the entire POPULATION is called a *parameter*. A statistic differs from a PARAMETER in at least two aspects. First, a statistic describes or summarizes a characteristic of the sample, but a parameter refers to a numerical summary of the population from which that sample was drawn. Note that values of sample statistics depend on the sample selected. Because their values vary from sample to sample, sample statistics are sometimes called RANDOM VARIABLES to reflect that their values vary according to the sample selected.

Second, the value of a parameter is usually unknown or not feasible to obtain, but the quantity of a statistic can be directly computed from the sample data. The primary focus of most social research is the parameter of the population, not the sample statistic calculated from selected observations. Thus, if the population of one's research interest is small and data exist for an entire population, then one would normally explore the population parameter rather than the sample statistic. However, in most social studies, researchers still use sample statistics to make inferences on the values of population parameters because (a) populations for social research are usually very large, so it is impractical to collect data for the entire population; and (b) one can still have good precision in making inferences about (or guessing) the true values of population parameters *if* appropriate sampling techniques and statistical methods are utilized. For example, samples taken by professional polling and research institutions, such as the Gallup, the General Social Survey, and the National Election Study, typically contain about 1,000 to 2,000 subjects.

One important issue for using a sample statistic to PREDICT or infer a parameter value is how accurate the prediction is. In other words, it is useful for data analysts to know the likely accuracy for predicting the value of a parameter by using a sample statistic. A conventional way to account for the uncertainty of predictions is to report the CONFIDENCE INTERVALS of POINT ESTIMATES of parameters. We usually use analogous sample statistics as the point estimates

(e.g., sample mean and population mean, sample variance and population variance). In addition, researchers may also set a “desired precision” of prediction before data collection. This desired precision (also defined as error bounds) in turn can help researchers decide the required sample size for obtaining reliable sample statistics. More information on confidence intervals of sample statistics and the relationship between sample size and accuracy of estimation using sample statistics may be found in William L. Hays (1994), Alan Agresti and Barbara Finlay (1997), and Richard A. Johnson and Gouri K. Bhattacharyya (2001).

—Cheng-Lung Wang

REFERENCES

- Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences*. Englewood Cliffs, NJ: Prentice Hall.
- Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.
- Johnson, R. A., & Bhattacharyya, G. K. (2001). *Statistics: Principles and methods* (4th ed.). New York: Wiley.

STATISTICAL COMPARISON

Most social scientists would agree that one of the *raison d'être* for the social sciences is comparison, which provides insight that we would not gain otherwise. The tradition of comparison goes back to the beginning of the social sciences and is evidenced in the works of Emile Durkheim and Max Weber. Today's social research of a comparative nature ranges from the COMPARATIVE RESEARCH of societies and nation-states, whether or not large national samples are used, to small-*N* case studies using the COMPARATIVE METHOD. As long as comparative researchers use a certain form of numerical data, they will typically make some kind of *statistical comparison*. A search on the Internet of “statistical comparison” would result in hundreds of hits that contain the term, some of which are social science studies. But what type of statistical comparison did the studies do? Whereas some are simply a listed comparison of two sets of figures, such as rates and AVERAGES, from two aggregates of people, others may involve some form of SIGNIFICANCE TESTING, with the aim to provide an INFERENCE about the POPULATION from SAMPLE statistics. The purpose of this entry is to give an overview of the wide range of possible statistical comparisons.

COMPARISONS OF MEANS

Often, researchers are interested to see if one group of people on average has certain attributes at a higher level than another group of people. This is a comparison of two MEANS. For example, one may be interested in comparing the mean incomes [$E(Y_a)$ and $E(Y_b)$] of the two groups of people *a* and *b*. The NULL HYPOTHESIS states that $E(Y_a) = E(Y_b)$, and typically, the *t*-TEST is applied when sample data are analyzed. A comparison like this lies at the heart of an EXPERIMENT. For further discussions along these lines, see Anderson et al. (1980).

Although the selection of subjects into the TREATMENT and CONTROL GROUPS is randomized in an experiment, in most social science research, the group variable in comparison is fixed (such as sex, race, and nation of origin), regardless of whether the individuals are drawn from one sample or separate samples.

The idea of comparing two means can be extended to making MULTIPLE COMPARISONS. The methodology for comparing multiple means includes many alternatives, such as the BONFERRONI TECHNIQUE, the SCHEFFÉ TEST, and Tukey's method. A major purpose of the one-factor ANALYSIS OF VARIANCE is also to assess potentially different response means from multiple groups.

COMPARISONS OF VARIANCES

If one is interested in assessing the difference in variance of two distributions, such as income, then one is concerned about the equality of the two variances; or, alternatively, the null hypothesis is $var(Y_a) = var(Y_b)$. The *F* TEST can be applied to the assessment of variance equality. This will be a useful test, especially when the two mean incomes may be identical but one income distribution may represent a perfect middle-class society where everyone has an income level located around the mean, whereas the other distribution comes from a society with much inequality where earning powers differ a great deal. For the analysis of variance inequality among multiple groups, LEVENE'S TEST is appropriate.

COMPARISONS OF DISTRIBUTIONS

Although variances are a property of DISTRIBUTIONS, there are more formal ways of comparing distributions, which can be studied in terms of their *proportionate* distribution, their *relative* distribution, and their *theoretical* distribution. All of these are

concerned about the shape and the general differences between two (or more) distributions as opposed to the differences in value between two (or more) parameters, such as the mean and the variance.

The method employing the Lorenz curve and the GINI COEFFICIENT showcases the first approach, where the researcher asks the question about the cumulative proportion of Y (or income) that is held by the cumulative population share. This method is popular for studying income inequality, and two (or more) overlaying Lorenz curves can directly compare two (or more) distributions with associated Gini coefficients to assist the comparison.

The RELATIVE DISTRIBUTION METHOD is concerned with full distribution comparison based on the so-called relative distribution. Simply defined, the relative distribution is a distribution resulting from computations involving the two distributions under investigation. Either the cumulative distribution function or the PROBABILITY DENSITY FUNCTION (PDF) of the relative distribution can be studied. For example, the PDF of the relative distribution of two income distributions is uniform.

Distributions can also be compared against a theoretical model. For example, researchers may be interested in whether both distributions follow the NORMAL DISTRIBUTION. It is possible that the distribution from one sample follows the normal and the distribution from another sample follows the log-normal. Such comparison may as well assist in identifying possible ASSUMPTION violations of a statistical method like the GENERAL LINEAR MODEL.

COMPARISONS OF RATES

Sometimes, researchers are interested in comparing rates of event occurrence instead of mean values of a variable. For example, a study may compare the unemployment rates of two populations. Because of the effect of potential CONFOUNDING factors, the two rates may not be directly comparable without some kind of ADJUSTMENT. A method of choice in this case is RATE STANDARDIZATION, which standardizes the rates in comparison with a standard population composition (i.e., ridding the rates of the confounding effect).

COMPARISONS OF REGRESSION PARAMETERS

When conducting a MULTIPLE REGRESSION ANALYSIS to study the impact of certain INDEPENDENT VARIABLES

on the DEPENDENT VARIABLE, researchers often are interested in comparing these parameters across samples or groups. For example, in studying income, one may be interested in comparing the effect of age (β_1) and of educational attainment (β_2) between men (m) and women (f). There are three possible null hypotheses:

1. $H_0 : \beta_{1m} = \beta_{1f}$
2. $H_0 : \beta_{2m} = \beta_{2f}$
3. $H_0 : \beta_{1m} = \beta_{1f} \text{ and } \beta_{2m} = \beta_{2f}$

There are several ways of testing Hypotheses 1 and 2. A popular way is to run a regression analysis on the combined sample with two interaction terms, one being the interaction between age and the sex DUMMY VARIABLE, and the other being the interaction between educational attainment and the sex dummy. A t -test of either of the interactions assesses either Hypothesis 1 or 2. Hypothesis 3 may also be assessed in different ways. One way is to run two regression models, one with the two interaction terms and the other without them; then, an F test of the R -SQUARED incremental is conducted. A variation of Hypothesis 3 tests not just the equality of the parameters for age and education but also that for the sex dummy variable. The method of assessment is the CHOW TEST, which evaluates whether the two sex groups differ at all in their income levels.

COMPARISONS IN THE GENERALIZED LINEAR MODEL

The hypotheses discussed above can be easily extended to other regression-type models, such as the LOGIT, the PROBIT, the MULTINOMIAL LOGIT, and the POISSON REGRESSION models, which are all members of the GENERALIZED LINEAR MODEL. Although the hypotheses can be formed similarly, the testing will be different. Interested readers are referred to Liao (2002).

COMPARISONS IN OTHER STATISTICAL MODELS

Similarly, comparisons can be made in more advanced methods such as LATENT CLASS ANALYSIS, STRUCTURAL EQUATION MODELING, and MULTILEVEL ANALYSIS. One's earning power can perhaps be better studied as a LATENT VARIABLE. Because individuals belonging in the same higher-level unit are more alike, they behave as DEPENDENT OBSERVATIONS, and the multilevel nature of the social structure can and should

be taken into account. Statistical comparison in these methods, as well as the methods covered so far, is discussed in greater detail in Liao (2002).

—Tim Futing Liao

REFERENCES

- Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandalae, W., & Weisberg, H. I. (1980). *Statistical methods for comparative studies: Techniques for bias reduction*. New York: Wiley.
- Liao, T. F. (2002). *Statistical group comparison*. New York: Wiley.

STATISTICAL CONTROL

Statistical control refers to the technique of separating out the effect of one particular INDEPENDENT VARIABLE from the effects of the remaining variables on the DEPENDENT VARIABLE in a MULTIVARIATE ANALYSIS. This technique is of utmost importance for making valid INFERENCES in a statistical analysis because one has to hold other variables constant in order to establish that a specific effect is due to one particular independent variable. In the statistical literature, phrases such as “holding constant,” “controlling for,” “accounting for,” or “correcting for the influence of” are often used interchangeably to refer to the technique of statistical control. Statistical control helps us to better understand the effects of an independent variable, say, X_1 , on the dependent variable Y because the influence of possibly CONFOUNDING or SUPPRESSING variables X_j ($j > 1$) is separated out by holding them constant.

How is this technically done when we estimate a particular MODEL? Suppose we find that campaign spending (X_1) increases turnout (Y) at the electoral district level. In order to establish the campaign spending effect, we want to know whether this relationship is confounded, suppressed, or SPURIOUS if we include control variables in our model that account for alternative explanations. The expected closeness of a district race (X_2) is such an alternative explanation. We expect that close races should particularly motivate voters to turn out on Election Day. Consider the following LINEAR regression model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

The specific effect β_1 of campaign spending on turnout in the classical REGRESSION framework is estimated as

$$\hat{\beta}_1 = \frac{\sum (X_1 - \hat{X}_1)(Y - \hat{Y})}{\sum (X_1 - \hat{X}_1)^2},$$

where $\hat{X}_1 = \alpha_0 + \alpha_1 X_2$ is predicted from a regression of X_1 on X_2 ; and $\hat{Y} = \gamma_1 + \gamma_2 X_2$ is predicted from a regression of Y on X_2 . Thus, the estimated campaign spending effect $\hat{\beta}_1$ solely depends on variation in X_1 and Y , because any linear impact of X_2 in the terms $X_1 - \hat{X}_1$ and $Y - \hat{Y}$ is subtracted. Consequently, $\hat{\beta}_1$ is calculated based on terms that are independent of X_2 . When estimating the specific effect of X_1 on Y , we hold constant any possible distortion due to X_2 because the relationship between X_1 and Y no longer depends on X_2 . Thus, statistical control separates out the specific effect $\hat{\beta}_1$ of X_1 on Y corrected for any influence of X_2 . At the same time, statistical control also disentangles the specific effect $\hat{\beta}_2$ of X_2 on Y corrected for any influence of X_1 . Thus, this technique does not treat our key explanatory variable any different from other (control) variables. The logic behind statistical control can be extended to more independent variables even for GENERALIZED LINEAR MODELS.

Statistical control comes with simplifying ASSUMPTIONS, though. Ideally, one likes to implicitly hold confounding factors constant, such as in experimental control through randomization. Because this is often not possible with observational data, one has to explicitly identify factors before one can statistically control for them. Furthermore, statistical control assumes that possible confounding effects are linear and additive, that is, they are independent of the particular level of any other variable in the model. Coming back to the running example, the expected closeness of the race is assumed to have the same impact on campaign spending and turnout no matter how close the race is expected to be. If these assumptions prove to be wrong, one can include product terms of the variables in the model to control for possible conditional or INTERACTION EFFECTS or restrict estimation of the model to subpopulations in which the assumptions hold.

—Thomas Gschwend

REFERENCES

- Gujarati, D. N. (2003). *Basic econometrics* (4th ed.). Boston: McGraw-Hill.

- Lewis-Beck, M. S. (1995). *Data analysis: An introduction*. Thousand Oaks, CA: Sage.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

STATISTICAL INFERENCE

Statistical inference is the process of making PROBABILITY statements about POPULATION PARAMETERS when we have only SAMPLE statistics. Population parameters are fixed quantities that are rarely known with certainty. If we know the population, then we can calculate (rather than estimate) population parameters. However, we usually do not know the population, so we must estimate population parameters from a sample. Such PARAMETER ESTIMATES always have some associated uncertainty because they are subject to SAMPLING ERROR. Therefore, we make statistical inferences about their likelihood of reflecting the true population parameter.

Understanding statistical inference requires understanding the concept of a SAMPLING DISTRIBUTION. A sampling distribution is a probability distribution of outcomes on a sample statistic when drawing all possible random samples of size N from a population. Often, a sampling distribution is represented as a two-dimensional graph, with each possible value of the sample statistic arrayed on the horizontal axis and the probability of each possible value of the sample statistic arrayed on the vertical axis. In reality, social science researchers typically draw only a single sample from the sampling distribution. However, knowledge of the sampling distribution enables us to make probability statements about sample statistics relative to the true population parameter.

We rely on fundamental theorems of statistics to make probability statements about whether a sample statistic is reflective of the true population parameter. For example, a fairly weak tool of statistical inference is Chebychev's theorem, which states that the probability of an estimate lying more than k standard deviations away from the mean of its distribution is $\frac{1}{k^2}$. For example, suppose we sample randomly from a distribution of any shape and obtain a POINT ESTIMATE of a population parameter. Then, Chebychev's theorem would state that the probability that the point estimate is more than 2 STANDARD DEVIATIONS away from the true mean of the SAMPLING DISTRIBUTION of estimated parameters is only $\frac{1}{4}$.

A more useful set of theorems enabling statistical inference is the family of central limit theorems. The Lindberg-Levy variant of the CENTRAL LIMIT THEOREM states that if repeated samples of size N are drawn from a probability distribution of any shape with mean μ and variance σ^2 , then as N grows large, the sampling distribution of sample means approaches normality with mean equal to the population mean μ and variance equal to $\frac{\sigma^2}{N}$. This result is important because it implies that regardless of the distribution of the population from which a sample statistic is drawn, the sampling distribution of sample statistics will be normally distributed with mean equal to the true population parameter and variance computable from the population variance and N . With this information, probability statements can be made based on the sample statistic and its variance.

A wide variety of other distributions can be used in statistical inference depending on the nature of the statistic to be inferred and the assumptions the analyst is willing to make. For example, it may be unrealistic to treat the population variance as known. In this case, the t -distribution is often used. Other distributions that are commonly used are the CHI-SQUARED and F distributions. The chi-squared distribution is commonly used in situations involving the sums of squares of normally distributed variables. The F distribution is often used for evaluating the ratio of independent chi-square variables.

—B. Dan Wood and Sung Ho Park

REFERENCES

- Blalock, H. M., Jr. (1979). *Social statistics*. New York: McGraw-Hill.
- Casella, G., & Berger, R. L. (2001). *Statistical inference* (2nd ed.). New York: Wadsworth.
- Judge, G. G., Hill, R. C., Griffiths, W. E., Lutkepohl, H., & Lee, T. (1988). *Introduction to the theory and practice of econometrics*. New York: Wiley.
- Spiegel, M. R., Schiller, J., & Srinivasan, R. A. (2000). *Schaum's outline of probability and statistics* (2nd ed.). New York: McGraw-Hill.

STATISTICAL INTERACTION

For mathematical statisticians, statistical interaction refers to those situations in MULTIVARIATE ANALYSES where the relationships among variables differ

depending on different values of the variables. In the simplest three-variable case, the relationship between two variables *A* and *B* is different depending on the values of a third variable *C*. If there is no interaction, relationships among variables are *additive*—that is, for the case of a single DEPENDENT VARIABLE, we can sum the effects of INDEPENDENT VARIABLES to determine the overall effect of all the independent variables on the dependent variable, and this sum will be constant, regardless of the values of the different independent variables. Statistical relationships can be NONADDITIVE, as represented by interaction effects. Physicists refer to this as *superposition*. If there is interaction, then superposition breaks down. Such failure of superposition is a characteristic of NONLINEAR systems.

There are a variety of statistical procedures that can detect interaction. ANALYSIS OF VARIANCE (ANOVA) techniques typically allow for the specification of models that include interaction terms, and there are exploratory analogues to ANOVA techniques. There are two variants on reporting interaction. One is to write an interaction term into a linear equation. The other is to report a model of determination that specifies that interaction is significant. Sometimes, we can specify a social process that engenders the interaction. For example, in a study of the relationship between housing and health conditions, it was found that for all age groups under 65, there was a positive relationship between quality of housing and quality of health. For those over 65, the relationship was negative. The causal process was that poor health was a determinant of the allocation of elderly households to very high-quality sheltered accommodation. In general, housing conditions had a causal influence on health. For the elderly, health had a causal influence on housing conditions (Byrne, McCarthy, Harrison, & Keithley, 1986). However, such straightforward explanations are not always possible.

Marsh (1982) noted that “interaction is a headache from a technical point of view but most exciting from the standpoint of substantive sociology” (p. 91). From a REALIST perspective, the existence of interaction is a challenge to the whole analytical program that informs traditional statistical reasoning. If we use a language of VARIABLES, at the very least the existence of interaction demonstrates that causal processes are local/contextual rather than universal. In other words, it is impossible to specify general laws because “circumstance [values of interacting variables] alter

cases.” Another way of putting this is to say that we are dealing with complex and contingent generative mechanisms. However, interaction might also indicate the presence of *emergence*—the scientific term that summarizes the expression “the whole is greater than the sum of its parts.” If we are dealing with *complex systems* with emergent properties, then there is an argument to the effect that VARIABLES are reifications rather than realities, and we should address the properties of whole systems rather than of artificial and unreal abstractions from them. At the very least, the existence of interaction alerts us to the possibility of complex relationships, and we might argue that the detection of interaction in a multivariate analysis invalidates any “linear” approach to reasoning about cause in the system(s) being investigated.

—David Byrne

REFERENCES

- Byrne, D. (2002). *Interpreting quantitative data*. London: Sage.
 Byrne, D. S., McCarthy, P., Harrison, S., & Keithley, J. (1986). *Housing and health*. Aldershot, UK: Gower.
 Marsh, C. (1982). *The survey method*. London: Allen and Unwin.

STATISTICAL PACKAGE

Statistical packages are collections of software designed to aid in statistical analysis and data exploration. The vast majority of quantitative and statistical analysis relies upon statistical packages for its execution. An understanding of statistical packages is essential to correct and efficient application of many quantitative and statistical methods.

Although researchers can, and sometimes do, implement statistical analyses in generalized programming languages such as FORTRAN, C++, and Java, statistical packages offer a number of potential advantages over such “hand coding.” First, statistical packages save the researcher time and effort: Statistical packages provide software for a range of analyses that would take years for an individual programmer to reimplement. Second, statistical packages can provide a unified operating framework and common interface for data manipulation, visualization, and statistical analysis. At their best, tools within a package can easily be combined and extended, enabling the

researcher to build upon previous work. Furthermore, even where a statistical procedure is simple enough or specialized enough to be programmed by hand, statistical packages offer user interfaces that make it far easier for others to use new procedures. Third, statistical packages are more likely to produce correct results than are hand-coded routines. Even simple tasks, such as computing the standard deviation, can be tricky to implement in a way that yields reliably accurate results. Statistical packages are likely to be tested more extensively for bugs and numerical problems, and the implementations of these packages are usually more familiar with modern numerical methods than is the typical researcher. Fourth, statistical packages are likely to be faster than average implementations: A considerable amount of effort and refinement goes into the choice of algorithm for a particular analysis or statistical tool, and also into tuning the software to work well on a specific computer architecture.

The chief problems researchers face when using statistical packages are choosing a package from the hundreds now available, and ensuring that the results obtained from it are correct and replicable. Obviously, each package will differ in terms of the features that it supports, although there is considerable overlap among many large packages. Commercial giants such as SAS[®] and SPSS[®], and large open-source packages such as “R,” offer thousands of data manipulation, visualization, and statistical features. The interfaces of the packages vary considerably as well: from the point-and-click style of SPSS to the command-line-based approach of SAS and R. Unless a particular package offers a crucial feature, the researcher may want to consider a modern statistical language such as “S,” which has broad functionality and both commercial (“S+”) and open-source implementations (“R”) (see Venables & Ripley, 2002, for an introduction).

Less obviously, packages differ significantly in terms of speed, accuracy, and extensibility. When approaching a statistical package, researchers may want to ask: Has this package been tested for accuracy? Does it document the algorithms it uses to compute statistical analyses and other quantities of interest? Was it designed to efficiently process data of the size and structure to be used in the proposed research project? What programming facilities are available? Does the programming language support good programming practices, such as object-oriented design?

Are the internal procedures of the package available for inspection and extension? How easy is it to use this package with an external library, or as part of a larger programming and data analysis environment?

Although statistical packages are viewed as essential tools for constructing an analysis, they are often considered highly interchangeable. Thus, once results are obtained, researchers often fail to document the package used in subsequent publications. In fact, however, statistical packages vary greatly with respect to accuracy and reliability, and reported results may be dependent on the specific package and version, and the computational options used during the analysis (McCullough & Vinod, 1997).

There are several approaches to ensuring accurate results. One may want to replicate the results across different packages while using different algorithms, to check that different methods of computing the results yield the same answer. In packages that support extended precision, increasing the numerical precision of computations during the analysis can be useful in avoiding certain types of numerical error. Another approach is sensitivity analysis, where small amounts of noise are introduced into calculations, data, and/or starting values in order to verify that the solution found is numerically robust (Altman, Gill, & McDonald, 2003).

Finally, three steps should be taken to increase the replicability of the results. Because complex analysis may depend on the particular computational strategy used both in researchers’ analysis and within the statistical package, reporting (or archiving) the code used to run the analysis, the package and version used, and any computational options used during the runs is essential to ensuring replicable results. Archiving a copy of the statistical package itself may be warranted in some cases where exact replication is vital; although a few packages have the ability to replicate the syntax rules and computational options used in older versions, most packages change both the programming syntax and implementation over time. Second, many social science data sets are difficult or impossible to re-create—publicly archiving any unique data used for an analysis is often a necessity for successful replications. Third, the formats that statistical packages use are often both proprietary and subject to change over time, posing grave obstacles to later use. Researchers can ensure that the data remain accessible even as proprietary formats change by exporting data using a plain-text format, along with adequate documentation

of missing values, weights, and other special coding of the data (Altman et al., 2003).

—Micah Altman

REFERENCES

- Altman, M., Gill, J., & McDonald, M. P. (2003). *Numerical methods in statistical computing for the social sciences*. New York: Wiley.
- Gentle, J. (2003). *Computational statistics*. New York: Springer-Verlag.
- McCullough, B. D., & Vinod, H. D. (1997). The numerical reliability of econometric software. *Journal of Economic Literature*, 37(2), 633–665.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer-Verlag.

STATISTICAL POWER

In using SAMPLE DATA to test HYPOTHESES about POPULATION values, one possibility is that the data will lead you to retain the NULL HYPOTHESIS even though it is false. This is TYPE II error (β). To accentuate the positive rather than the negative, however, the term *power* has been used to define the converse of Type II error: The power of a statistical test is the probability of correctly rejecting the null hypothesis when it is false, or $P = 1 - \beta$. Statistical power is an extremely important concept because most social science research is designed to demonstrate that a particular variable does, in fact, influence behavior. We set out to find evidence that the “hypothesis of no effect” (the null hypothesis) is indeed false. We want to demonstrate that a particular form of therapy is effective, that an educational program works, that world events do influence political opinion, and so forth. To do these things, we desire high levels of statistical power.

Although statistical power will decrease when α , the probability of a TYPE I ERROR (rejecting a true null hypothesis), is set low, there are some principles of good EXPERIMENTAL DESIGN that can be used to increase power: use large samples, develop reliable response measures, select homogeneous subject populations, use repeated measures on the same subjects, and use STANDARDIZED and unambiguous tasks and instructions. One factor that we cannot control, however, is the true value of the POPULATION PARAMETER being estimated by our statistical test. The greater the discrepancy between the value stated in the

null hypothesis and the true situation, the greater the statistical power. Logically (and fortunately), we will be more apt to detect the effect of a strong (*powerful*) treatment than a weak one. Some researchers plot power functions or power curves to show how the power of their statistical test varies depending on the degree of falsity of the null hypothesis (Cohen, 1988). Conversely, some researchers will consult previous studies or conduct their own PILOT work to estimate the size of their experimental effect and then use this to determine the sample size they would need in their main study to attain a predetermined level of statistical power, say .90. (See Cohen, 1988, for how this is done.)

Hays (1994) likens statistical power to the power of a microscope. Just as a high-powered microscope increases our chances of detecting something not readily visible with a low-powered instrument, so does a high-powered test increase our chances of detecting when the null hypothesis is false. However, even a high-powered microscope is limited by the size of the object it is attempting to detect. So it is that, all other things being equal, the larger the effect that we are trying to detect in our research, the greater will be the power of our tests for detecting it. Good research techniques will also increase the likelihood of detecting the effect.

—Irwin P. Levin

REFERENCES

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.
- Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.

STATISTICAL SIGNIFICANCE

When an ESTIMATE achieves statistical significance, the thing estimated is likely to be not zero, and so becomes especially worthy of further attention. Significance tests are applied to UNIVARIATE statistics, such as the mean; to BIVARIATE statistics, such as PEARSON'S CORRELATION COEFFICIENT; and to multivariate statistics, such as the slope coefficient in a MULTIPLE REGRESSION equation. For example, if a Pearson's correlation coefficient between variable *X* and variable *Y* attains statistical significance at .05, the level most commonly used, then we can reject the

NULL HYPOTHESIS that X and Y are not related in the population under study. If, however, the correlation is not found significant at .05, we tend to think, at least initially, that the link between X and Y does not merit further study. Thus, statistical significance serves as a gatekeeper, encouraging further consideration of coefficients that “pass through the gate” and discouraging further consideration of coefficients that fail to “pass through the gate.”

—Michael S. Lewis-Beck

See also SIGNIFICANCE TESTING

STEM-AND-LEAF DISPLAY

Developed by John Tukey, a stem-and-leaf plot is a way of visually representing a DISTRIBUTION of data values. The stem represents a major unit for grouping the values, and the leaf represents a subdivision of that unit. Each possible stem value is marked even when no cases belong to that group. It then becomes possible to observe where values bunch, gaps in the distribution, and whether the distribution appears to be SKEWED. The stem-and-leaf plot can also indicate outlying values. In order to best represent the distribution of the data, the unit for the stem has to be selected so as not to condense the data too much, but also so as not to provide so much detail that the pattern of the data becomes indistinguishable. Stem values can be divided or multiplied by 2, 5, or 10 to expand or condense the summary of the distribution provided. For example, Table 1 shows a set of 30 household income values from a British survey of the late 1990s.

If the stem values are set in units of 10, the ensuing stem-and-leaf values from the original data values are those shown in the second column of the table, and they are represented on the stem-and-leaf plot as illustrated in Figure 1.

Here, in what can be described as a form of horizontal HISTOGRAM, we can see that the distribution peaks in the £130–£139 range, although it also has a smaller, earlier peak at incomes around £80. We can see that there are gaps in the 110s, the 160s, and the 190s. These could represent breaks in the distribution that indicate separate distributions, although they could also stem from the small size of the sample. There are no values above £200 up to the value of £274, so rather than represent the empty stem points and thus

Table 1 Sample of 30 Cases From a British Income Distribution and Values as Marked on Stem-and-Leaf Plots

<i>Income (£)</i>	<i>Value where points on stem = 10</i>	<i>Value where points on stem = 20</i>	<i>Value where points on stem = 50</i>
61	6 1	0s 6	0. 6
63	6 3	0s 6	0. 6
74	7 4	0s 7	0. 7
74	7 4	0s 7	0. 7
80	8 0	0. 8	0. 8
84	8 4	0. 8	0. 8
85	8 5	0. 8	0. 8
86	8 6	0. 8	0. 8
88	8 8	0. 8	0. 8
95	9 5	0. 9	0. 9
99	9 9	0. 9	0. 9
107	10 7	1* 0	1* 0
120	12 0	1t 2	1* 2
121	12 1	1t 2	1* 2
122	12 2	1t 2	1* 2
123	12 3	1t 2	1* 2
131	13 1	1t 3	1* 3
133	13 3	1t 3	1* 3
133	13 3	1t 3	1* 3
135	13 5	1t 3	1* 3
136	13 6	1t 3	1* 3
139	13 9	1t 3	1* 3
142	14 2	1f 4	1* 4
144	14 4	1f 4	1* 4
150	15 0	1f 5	1. 5
156	15 6	1f 5	1. 5
171	17 1	1s 7	1. 7
188	18 8	1. 8	1. 8
200	20 0	2* 0	2* 0
274	27 4	2s 7	2. 7

creating a longer and more straggling stem, the top value of £274 has simply been indicated at the end as an OUTLIER.

In this stem-and-leaf plot, the exact values of the original data have been retained: We know that the lowest value is £61 and the third highest is £188. If the stem is reduced from divisions of 10, the values may be simplified. For example, if each point on the stem represents a range of £20, then the leaf can simply distinguish points according to which of the two £10 ranges they fall into. In Figure 2, the values from

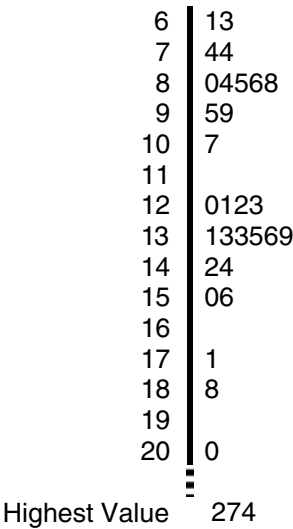


Figure 1 Stem-and-Leaf Plot of Data in Table 1

NOTE: Stem unit is £10; leaf unit is £1.

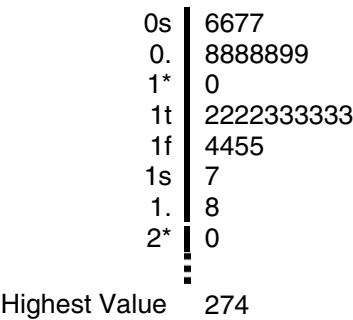


Figure 2 Stem-and-Leaf Plot of Data in Table 1

NOTE: Stem unit is £20; leaf unit covers the range £x0–£x9.

the third column of Table 1 have been illustrated on a stem-and-leaf plot. Here, the number of the stem point gives the number of hundreds, with the following symbol indicating which £20 division is referred to: The “*” represents the range from 0 to 19, “t” the twenties and thirties, “f” the forties and fifties, “s” the sixties and seventies, and “.” the eighties and nineties. The leaf number simply indicates which of the two possibilities is actually contained within the data.

Here, the shape of the distribution with two clear peaks and a trailing end is emphasized, although at the loss of more detailed definition. In this example, if the scale is reduced even further, such as into £50 stem

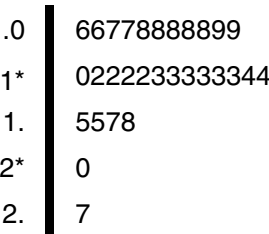


Figure 3 Stem-and-Leaf Plot of Data in Table 1

NOTE: Stem unit is £50; leaf unit covers the range £x0–£x9.

divisions, as illustrated in Figure 3, the stem-and-leaf plot ceases to become so informative.

—Lucinda Platt

REFERENCES

Marsh, C. (1988). *Exploring data: An introduction to data analysis for social scientists*. Cambridge, UK: Polity.
Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

STEP FUNCTION. See INTERVENTION ANALYSIS

STEPWISE REGRESSION

Given a set of potential EXPLANATORY VARIABLES that could be included in a LINEAR REGRESSION equation, and little or no theoretical and subject knowledge basis for choosing one variable over another, automated methods have been developed for choosing which variables to include in the regression. Stepwise regression is a class of model-building ALGORITHMS that incrementally includes or excludes variables based on how much they contribute to fitting the dependent variable of the regression. Although data-driven automatic MODEL selection procedures are available in most statistical software packages and may seem attractive, a model chosen by the stepwise procedure will, in general, lack optimality properties and fail to provide correct statistical inference. A stepwise-selected model generally will not be the correct model, nor the

best fitting model, nor the model that provides good out-of-sample predictions.

There are a variety of stepwise algorithms, but each follows the same basic logic. The stepwise algorithm incrementally chooses whether to include a variable or set of variables into a regression based on the variable's contribution to a particular fit criterion. The most common criterion for linear additive models is the minimization of the residual sum of squares (RSS) at each step. More general approaches have been developed with a similar logic based on log-likelihood statistics for use with CATEGORICAL and other types of data. The search for a "better" model is stopped, and the final model is thus chosen when a predetermined maximum number of variables have been selected, or the fit of the model is changed by less than a predetermined threshold when including or excluding additional variables.

The two most common implementations are forward selection and backward selection. Forward selection is more often used when the number of potential regressors (k) is close to or exceeds the number of observations in the data set (n). In forward selection, an initial model is usually fit with only the mean,

$$Y_i = \beta_0 + e_i$$

and at each subsequent p th step, one variable $x_{(p)}$ from among the remaining explanatory potential regressors ($x_1, x_2, \dots, x_k$) is added,

$$y_i = \beta_0 + \beta_1 x_{(1)i} + e_i$$

$$\dots$$

$$y_i = \beta_0 + \beta_1 x_{(1)i} + \beta_2 x_{(2)i} + \dots + \beta_p x_{(p)i} + e_i.$$

Using LEAST SQUARES to estimate the parameters, $\hat{\beta}_j$, the RESIDUAL SUM OF SQUARES with p variables is

$$RSS_p = \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \sum_{j=1}^p \hat{\beta}_j x_{(j)1} \right)^2.$$

At the p th step, each of the $k - p - 1$ remaining excluded variables is added to the current model individually, and the RSS from these $k - p - 1$ regressions is retained. The variable that is added to the regression equation is the one that induces the greatest incremental change in the residual sum of squares,

$$F = \frac{RSS_{p-1} - RSS_p}{RSS_p / (n - p - 2)}.$$

Descriptions of stepwise regression may alternatively state that the procedure uses one of the following

equivalent ranking criteria, selecting a variable that (a) increases R^2 the most, (b) has the largest t -statistic, or (c) has the largest partial correlation with the dependent variable. Variables are usually added until either there are no more variables with a significant F statistic that remain to be added, or a user-selected upper bound on the number of variables (p) is reached. Note that in the context of a stepwise search, the F statistic is not actually drawn from an F DISTRIBUTION.

Backward selection does essentially the reverse of forward selection, starting with a full set of variables included in the regression, and then sequentially eliminating the variable that, equivalently, (a) decreases R^2 the least, (b) has the smallest F or t -statistic, or (c) has the smallest partial correlation with the dependent variable. Once eliminated, a variable is not subsequently reconsidered for inclusion in the model. Note that if the number of observations is less than the number of variables, including all variables in a regression is not feasible. In the case of $n < k$, either the backward selection procedure must be abandoned or the researcher must make an initial choice of eliminating at least $k - n - 1$ variables prior to beginning the algorithm.

Stepwise regression fails as a tool for model selection for both theoretical and practical reasons. With respect to the final model that is chosen by stepwise regression, the method invalidates much of the standard theory for statistical inference when the same data are used for model building/selection and HYPOTHESIS TESTING (Miller, 1990). In particular, the parameter estimates on the selected variables will be biased upwards in magnitude because of the SELECTION BIAS of the procedure. Inference based on the selected model is also incorrect because the CONFIDENCE INTERVALS on coefficients will be too small, and F tests on subsets of parameters will be artificially significant because the variance of the residuals will be underestimated. Moreover, not only will stepwise regression generally fail to find the correct model that generated the data, but stepwise regression is not even likely to find the best fitting model! The problems associated with stepwise regression are minimized in the rare case where all the potential regressors are orthogonal to one another, and the problems are exacerbated as the CORRELATION between variables increases.

There are alternatives to relying on stepwise regression to sift through large numbers of competing explanatory variables. If multiple explanatory variables are believed to be manifestations of a small

number of common latent measures, then it is better to include summaries of these latent measures as explanatory variables in the regression. For example, it is possible to reduce the dimensionality of multiple explanatory variables by using only their PRINCIPAL COMPONENTS as REGRESSORS. If groups of explanatory variables exist that each represent competing theories, then it is better to formulate non-nested hypothesis tests to directly assess the significance of the difference between the competing theories. It is better to let theory and substantive knowledge of the data guide model building instead of relying on generally misleading automated procedures such as stepwise regression.

—Jonathan Wand

REFERENCE

Miller, A. J. (1990). *Subset selection in regression*. New York: Chapman Hall.

STOCHASTIC

Variables whose values cannot be fully controlled or determined are stochastic. If a variable measures the outcomes of N trials of an experiment, it is a stochastic variable if the outcomes cannot be entirely determined or predicted prior to observation. The random variation in outcomes is due to chance, and hence, stochastic variables are defined with probability distributions rather than with exact or deterministic functions. More generally, a stochastic dependent variable, $y = g(\mathbf{X}) + \varepsilon$, is composed of a deterministic function, $g(\mathbf{X})$, which could be simply a constant, and a stochastic disturbance or ERROR term, ε .

SIGNIFICANCE

Stochastic is a fundamental concept in social science methodology because all causal processes (dependent variables) are assumed to be stochastic. In the natural sciences, causal relationships are exact rather than stochastic, so researchers can calculate the PARAMETERS that define a relationship by constructing a controlled experiment and running trials that cover the full range of values taken by the explanatory variables. When the causal relationship is exact, trials of the experiment are points on the deterministic function rather than a sample drawn from the PROBABILITY DISTRIBUTION centered on that function, because there

is no SAMPLING ERROR if the dependent variable is not stochastic. In contrast, researchers in the social sciences must use QUASI-EXPERIMENTS, face greater limitations in MEASUREMENT, and investigate more complex causal processes involving human behavior rather than the interaction of physical elements. In turn, all causal processes in the social sciences are assumed to be stochastic in nature and hence must be investigated through SAMPLING and STATISTICAL INFERENCE. Consider the implication of this assumption for predicting values of a stochastic dependent variable, $y = g(\mathbf{X}) + \varepsilon$. Even if the deterministic part, $g(\mathbf{X})$, is known, prediction is still made with ERROR because the stochastic component, ε , causes the values of the dependent variable to vary around the conditional mean of y given $\mathbf{X}$, defined by $g(\mathbf{X})$.

The assumption that all dependent variables are stochastic is generally justified in three ways. First, the stochastic disturbance accounts for the omission of innumerable minor, idiosyncratic factors. The net effect of these omitted factors on the dependent variable is assumed to vary randomly across observations. Second, dependent variables in the social sciences are generally, if not always, measured with ERROR. Third, human behavior is inherently RANDOM. Factors ranging from incomplete information to strategic behavior complicate decision making in economic, political, and social contexts and thereby prevent people from behaving in a purely deterministic manner. This justification is often the basis for arguing that all populations in the social sciences are infinite, because even history is a sample of all possible histories if human behavior is stochastic.

Explanatory variables are often assumed to be nonstochastic or fixed in repeated samples. When explanatory variables are stochastic, the crucial issue is whether they are correlated with the disturbance. If correlated, they produce BIASED PARAMETER ESTIMATES. If uncorrelated, the explanatory variables represent ancillary REGRESSORS (i.e., the marginal probability distributions of the regressors do not depend on the parameters of interest). In the context of dynamic models, this condition is labeled weak EXOGENEITY. When stochastic regressors are ancillary or weakly exogenous, conditioning on their observed values allows them to be treated as fixed.

HISTORY

The word *stochastic* comes from the Greek word *stokhos*, meaning “a bull’s eye” or target. The process

of throwing darts produces stochastic outcomes that could be modeled as a probability distribution with the bull's eye at the center. Formal theorizing about stochastic outcomes began with a series of exchanges between Blaise Pascal and Pierre Fermat in 1654 about the game of dice, which established some fundamental concepts in PROBABILITY theory. Development of the theory of stochastic processes, however, did not begin until the early 1900s, when Andrei Markov proposed the concept of the MARKOV CHAIN. The Markov chain is a sequence of random variables in which the value (state) of the future variable is determined by the value (state) of the present variable but is independent of the process by which the value (state) of the present variable arose from the values (states) of previous variables in the sequence. Extension of this concept from a discrete to a continuous state space (set of possible variable values) defines a general stochastic process (also called a Markov process).

APPLIED EXAMPLES

Consider the random event of flipping a coin, where H and T denote the two possible outcomes. Given a fair coin, the two events are equally likely, $P(H) = P(T) = 0.5$. Yet knowing the probability distribution for the outcomes of this event does not enable one to predict the outcome of any trial with certainty. Suppose p_N is the fraction of N trials for which the outcome is H . For $N = 6$, the likelihood that p_N equals the true parameter, $\pi = 0.5$, is only 31.25%. The stochastic nature of this event causes p_N to vary around its true value, π , thereby preventing calculation of the parameter and requiring the use of statistical inference.

Consider another example in the context of REGRESSION analysis. Suppose the causal process of interest is modeled as a linear function of a set of exogenous variables and a stochastic disturbance, $y = \mathbf{X}\beta + \varepsilon$, where $\mathbf{X}$ is the matrix of exogenous variables, β is the vector of unknown parameters, and ε is the disturbance. The deterministic part of the relationship is $\mathbf{X}\beta$, whereas ε represents the stochastic part. Regression analysis uses sample data to estimate β , which involves decomposing the variation of y into its explained part attributable to $\mathbf{X}$ (i.e., variation of the deterministic function) and its unexplained part attributable to ε (i.e., variation of the stochastic disturbance). As functions of the dependent variable, the estimator of β is a stochastic variable with a SAMPLING DISTRIBUTION, and predictions about the value of y given $\mathbf{X}$ are made with

ERROR due to uncertainty attributable to the stochastic disturbance as well as the joint estimation of the parameters. In order to simplify statistical inference, assumptions are often imposed on the probability distribution of the stochastic disturbance as part of ORDINARY LEAST SQUARES estimation.

—Harvey D. Palmer

REFERENCES

- Kennedy, P. (1998). *A guide to econometrics* (4th ed.). Cambridge: MIT Press.
 Zaman, A. (1996). *Statistical foundations for econometric techniques*. New York: Academic Press.

STRATIFIED SAMPLE

Often, SURVEY samples are designed in a way that takes into account knowledge about the POPULATION structure. This may involve an attempt to ensure that the SAMPLE has the same structure as the population in important respects or an attempt purposively to deviate from the population structure in order to increase the representation of certain subgroups. Such sample designs are referred to as *stratified sampling*, and the outcome of implementing the design is a *stratified sample*.

There are two types of stratified sampling: proportionate and disproportionate. Proportionate stratified sampling involves controlling the sample proportions in each stratum to equal the population proportions. If the strata are correlated with the survey measures, this will have the effect of increasing the precision of survey estimates (see DESIGN EFFECTS). Proportionate stratification can be achieved by either creating explicit strata and sampling independently from each, or sorting the SAMPLING FRAME units into a meaningful order and then sampling systematically from the ordered list (see SYSTEMATIC SAMPLING). These methods are known, respectively, as “explicit stratified sampling” and “implicit stratified sampling.” Proportionate stratification cannot have an adverse effect on the precision of estimates. The worst scenario is that strata turn out to have zero correlation with a particular survey measure, in which case stratification has no effect on precision and the sample is equivalent to a simple random sample (see SIMPLE RANDOM SAMPLING). Selection probabilities are not affected by proportionate stratification, so no BIAS can be introduced by a poor choice of strata.

Disproportionate stratification involves applying different sampling fractions (see SAMPLING FRACTION) in different strata. The objective is often to increase the sample size of one or more important subgroups for which separate estimates are required. In this situation, disproportionate stratification typically reduces the precision of estimates relating to the total study population (see DESIGN EFFECTS), but increases the precision of estimates for the oversampled stratum or strata. Disproportionate stratification can, however, also be used to *increase* the precision of total population estimates in situations where inherent variability differs greatly between the strata and can be estimated in advance. In such situations, precision is maximized by use of Neyman optimum allocation:

$$\frac{n_h}{N_h} \propto \frac{s_h}{\sqrt{C_h}}. \quad (1)$$

The optimum allocation rule states that the sampling fraction in stratum h , $\frac{n_h}{N_h}$, should be set proportional to the stratum standard deviation, s_h , and inversely proportional to the square root of the unit cost of data collection in the stratum, C_h .

In quantitative surveys, the choice of proportionate or disproportionate sampling is an important and explicit part of the design. In QUALITATIVE studies, there is no attempt to quantify the relationship between sample and population, so the distinction between proportionate and disproportionate sampling is not relevant. But a sample can still be said to be stratified if it has been designed deliberately to reflect certain aspects of population structure.

—Peter Lynn

REFERENCES

- Moser, C. A., & Kalton, G. (1971). *Survey methods in social investigation* (2nd ed.). Aldershot, UK: Gower.
- Scheaffer, R. L., Mendenhall, W., & Ott, L. (1990). *Elementary survey sampling*. Boston: PWS-Kent.
- Stuart, A. (1984). *The ideas of sampling*. London: Griffin.

STRENGTH OF ASSOCIATION

The strength of association shows how much two variables COVARY and the extent to which the INDEPENDENT VARIABLE affects the DEPENDENT VARIABLE.

In summarizing the relationship between two VARIABLES in a single summary STATISTIC, the strength of the association is shown by a value between 0

and 1. The larger the absolute value of the association measure, the greater the association between the variables. A value of 1 signifies a perfect association between the variables according to the association model that is being used. A value of 0 shows that there is a NULL relationship between the variables. Negative values for ordinal and numeric data occur when larger values of one variable correspond to smaller values of the other variable.

The strength of association between two variables corresponds to the TREATMENT difference in EXPERIMENTS. As such, it can be measured using ASYMMETRIC MEASURES, although it is more commonly measured with symmetric COEFFICIENTS whose value is the same regardless of which variable is the independent variable.

Many statistical measures of RELATIONSHIP have been proposed, and they differ in their assessment of strength of association. The “gold standard” is PEARSON’S R for numeric data, whose squared value shows the proportion of variance in one variable that can be accounted for by LINEAR prediction from the other variable. Several measures for non-numeric variables can similarly be interpreted as PROPORTIONAL-REDUCTION-IN-ERROR measures (Costner, 1965). Whereas some of these measures attain their maximum value of 1 when there is a strong MONOTONIC relationship between the variables, others (such as YULE’S Q) require only a weak monotonic relationship, as when larger values on one variable correspond to equal or larger values on the other variable.

Pearson’s r interprets statistical INDEPENDENCE as the condition for no relationship, as do such other coefficients as Goodman and Kruskal’s TAU and Yule’s Q . Weisberg (1974) distinguishes other null conditions for some coefficients. The LAMBDA COEFFICIENT for nominal variables, for example, has a value of 0 when the modal category on the dependent variable is the same for every category of the independent variable.

There are no strict criteria for evaluating whether an association is large or small, although relationships above .70 are usually considered large, and those below .30 are usually considered small.

The strength-of-association concept can be generalized to more than two variables. For example, a CANONICAL CORRELATION shows the strength of association between two sets of variables. The R -SQUARED statistic shows the strength of association represented by a REGRESSION EQUATION.

—Herbert F. Weisberg

REFERENCES

- Costner, H. L. (1965). Criteria for measures of association. *American Sociological Review*, 30, 341–353.
- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49, 732–764.
- Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53, 814–861.
- Weisberg, H. F. (1974). Models of statistical relationship. *American Political Science Review*, 68, 1638–1655.

STRUCTURAL COEFFICIENT

The term *structural coefficient* is used in discussions of PATH ANALYSIS and STRUCTURAL EQUATION MODELING (Duncan, 1975) to contrast it from the term *regression coefficient*, typically used in discussions of regression analysis. To motivate the term, it is often useful to represent the set of structural equations in the form of a *path diagram*. Figure 1 shows the path diagram for a student-level model of science achievement. This figure was used in Kaplan (2000) to motivate the discussion of path analysis. Path diagrams are especially useful pictorial devices because, if drawn accurately, then there is a one-to-one relationship between the diagram and the set of structural coefficients.

The system of structural equations representing the model in Figure 1 can be written as

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{B} \mathbf{y} + \boldsymbol{\Gamma} \mathbf{x} + \boldsymbol{\zeta}. \quad (1)$$

where $\mathbf{y}$ is a vector of observed endogenous variables, $\mathbf{x}$ is a vector of observed exogenous variables, $\boldsymbol{\alpha}$ is a vector of structural intercepts, $\mathbf{B}$ is a structural coefficient matrix that relates endogenous variables to each other, $\boldsymbol{\Gamma}$ is a structural coefficient matrix that relates endogenous variables to exogenous variables, and $\boldsymbol{\zeta}$ is a vector of disturbance terms where $\text{Var}(\boldsymbol{\zeta}) = \boldsymbol{\Psi}$ is the covariance matrix of the disturbance terms. Finally, let $\text{Var}(\mathbf{x}) = \boldsymbol{\Phi}$ be the covariance matrix for the exogenous variables.

The elements in $\mathbf{B}$ and $\boldsymbol{\Gamma}$ are the structural coefficients representing the relationships among the variables. The pattern of zero and nonzero elements in these matrices, in turn, are imposed by the underlying substantive theory. For example, in the model shown in Figure 1, a structural coefficient in $\mathbf{B}$ would be the path relating science achievement at grade 10 (SIGRA10) to the extent to which students find their work in the science class challenging (CHALLG). A

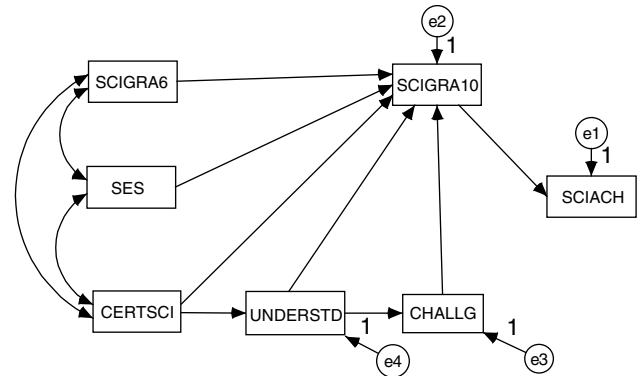


Figure 1 Education Indicators Model: Initial Specification

SOURCE: Kaplan (2000).

structural coefficient in $\boldsymbol{\Gamma}$ would be the path relating SIGRA10 to socioeconomic status (SES).

We can distinguish between three types of structural coefficients in $\mathbf{B}$ and $\boldsymbol{\Gamma}$. The first type are those that are to be estimated. These are often referred to as *free* parameters. Thus, in the model in Figure 1, the observable paths are the free parameters.

The second set of parameters is the parameters that are given a priori values that are held constant during estimation. These parameters are often referred to as *fixed* parameters. Most often, the fixed parameters are set equal to zero to represent the absence of the relationship. Again, considering the model in Figure 1, we theorize that there is no direct relationship between science achievement (SCIACH) and SES. Thus, the structural coefficient is fixed to zero. Finally, it is possible to constrain certain parameters to be equal to other parameters in the model. These elements are referred to as *constrained* parameters. An example of constrained parameters would be requiring that the effect of SCIGRA6 on SIGRA10 be the same as the effect of SES on SIGRA10.

Three additional features of structural coefficients are worth noting. First, as with regression coefficients in general, structural coefficients can be tested for significance. Second, when the structural coefficients are standardized, they are referred to as path coefficients (Duncan, 1975; Wright, 1934). Third, structural coefficients can be multiplied and added together to form coefficients of indirect and total effects (Duncan, 1975).

—David W. Kaplan

REFERENCES

- Duncan, O. D. (1975). *Introduction to structural equation models*. New York: Academic Press.
- Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.
- Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, 5, 161–215.

STRUCTURAL EQUATION MODELING

Structural equation modeling (SEM) can be defined as a class of methodologies that seeks to represent hypotheses about the means, variances, and covariances of observed data in terms of a smaller number of “structural” parameters defined by a hypothesized underlying conceptual or theoretical model. Historically, structural equation modeling derived from the hybrid of two separate statistical traditions. The first tradition is factor analysis, developed in the disciplines of psychology and psychometrics. The second tradition is simultaneous equation modeling, developed mainly in econometrics, but having an early history in the field of genetics and introduced to the field of sociology under the name PATH ANALYSIS. A more detailed discussion of the history of structural equation modeling can be found in Kaplan (2000).

THE GENERAL MODEL

The general structural equation model, as outlined by Jöreskog (1973), consists of two parts: (a) the structural part linking latent variables to each other via systems of simultaneous equations, and (b) the measurement part linking latent variables to observed variables via a restricted (confirmatory) factor model. The structural part of the model can be written as

$$\eta = \alpha + B\eta + \Gamma\xi + \zeta, \quad (1)$$

where η is a vector of endogenous (criterion) latent variables, α is a vector of structural intercepts, ξ is a vector of exogenous (predictor) latent variables, B is a matrix of regression coefficients relating the latent endogenous variables to each other, Γ is a matrix of regression coefficients relating endogenous variables to exogenous variables, and ζ is a vector of disturbance terms.

The latent variables are linked to observable variables via measurement equations for the endogenous variables and exogenous variables. These equations are defined as

$$y = \Lambda_y\eta + \epsilon, \quad (2)$$

and

$$x = \Lambda_x\xi + \delta, \quad (3)$$

where Λ_y and Λ_x are matrices of factor loadings, respectively, and ϵ and δ are vectors of uniqueness, respectively. In addition, the general model specifies variances and covariances for ξ , ζ , ϵ , and δ , denoted as Φ , Ψ , Θ_ϵ , and Θ_δ , respectively. A path diagram of a prototypical structural equation model is given in Figure 1.

Structural equation modeling is a methodology designed primarily to test substantive theories. As such, a theory might be sufficiently developed to suggest that certain constructs do not affect other constructs, that certain variables do not load on certain factors, and that certain disturbances and measurement errors do not covary. This implies that some of the elements of B , Γ , Λ_y , Λ_x , Φ , Ψ , Θ_ϵ , and Θ_δ are *fixed* to zero by hypothesis. The remaining parameters are *free* to be estimated. The pattern of fixed and free parameters implies a specific structure for the covariance matrix of the observed variables. In this way, structural equation modeling can be seen as a special case of a more general *covariance structure model* defined as $\Sigma = \Sigma(\Omega)$, where Σ is the population covariance matrix, and $\Sigma(\Omega)$ is a matrix-valued function of the *parameter vector* Ω containing all of the parameters of the model.

IDENTIFICATION OF STRUCTURAL EQUATION MODELS

A prerequisite for the estimation of the parameters of the model is to establish whether the parameters are *identified*. A definition of *identification* can be given by considering the problem from the covariance structure modeling perspective. The population covariance matrix Σ contains population variances and covariances that are assumed to follow a model characterized by a set of parameters contained in Ω . We know that the variances and covariances in Σ can be estimated by their sample counterparts in the sample covariance matrix S using straightforward formulae for the calculation of sample variances and covariances. Thus, the parameters in Σ are identified. What needs to be

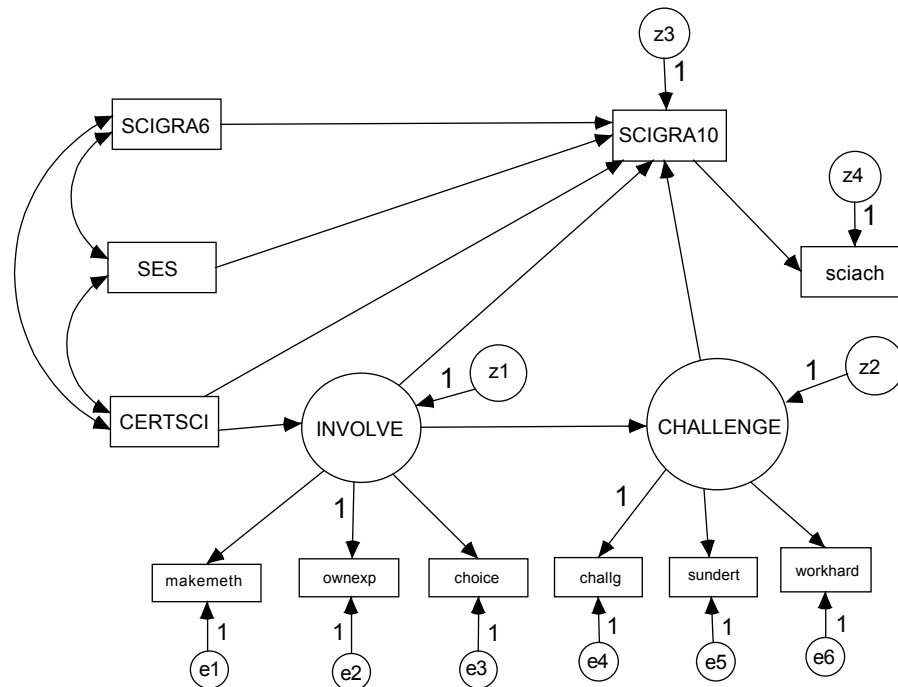


Figure 1 Structural Equation Model of Science Achievement Model
SOURCE: Kaplan (2000).

established prior to estimation is whether the unknown parameters in Ω are identified from the elements of Σ .

We say that the elements in Ω are *identified* if they can be expressed uniquely in terms of the elements of the covariance matrix Σ . If all elements in Ω are identified, then the model is identified. A variety of necessary and sufficient conditions for establishing the identification of the model exist and are described in Kaplan (2000); among these are the rank and order conditions of identification that came out of the econometric tradition.

ESTIMATION OF STRUCTURAL EQUATION MODELS

The problem of parameter estimation in structural equation modeling concerns obtaining estimates of the free parameters of the model, subject to the constraints imposed by the fixed parameters of the model. In other words, the goal is to apply a model to the sample covariance matrix S , which is an estimate of the population covariance matrix Σ that is assumed to follow a structural equation model. Estimates are obtained such that the discrepancy between the estimated covariance matrix $\hat{\Sigma} = \Sigma(\hat{\Omega})$ and the sample covariance matrix S is as small as possible.

Parameter ESTIMATION requires choosing among a variety of *fitting functions*. Fitting functions have the properties that (a) the value of the function is greater than or equal to 0; (b) the value of the function equals 0 if, and only if, the model fits the data perfectly; and (c) the function is continuous. Estimation can be conducted under the assumption of multivariate normality of the observed data such as MAXIMUM LIKELIHOOD ESTIMATION, or under more relaxed assumptions regarding the distribution of the data such as general weighted least squares estimation for continuous non-normal and/or categorical data.

TESTING

Once the parameters of the model have been estimated, the model can be tested for global fit to the data. The null hypothesis states that the specified model holds in the population. The corresponding alternative hypothesis is that the population covariance matrix is any symmetric (and positive definite) matrix. Under maximum likelihood estimation, the statistic for testing the null hypothesis that the model fits in the population is referred to as the LIKELIHOOD RATIO (LR) TEST. The large sample distribution of the likelihood ratio test is chi-square with degrees of freedom given by the

difference in the number of nonredundant elements in the population covariance matrix and the number of free parameters in the model.

In addition to a global test of whether the model holds exactly in the population, one can also test hypotheses regarding the individual fixed and freed parameters in the model. These include the Lagrange multiplier test (also referred to as the modification index) and the Wald test. The former examines whether freeing a fixed parameter will improve the fit of the model, whereas the latter examines whether fixing a free parameter will degrade the fit of the model. Strategies for the use of these methods have been described in Saris et al. (1987) and Kaplan (1990).

Alternative Methods of Model Fit

For more than 20 years, attention has focused on the development of alternative indexes that provide relatively different perspectives on the fit of structural equation models. The development of these indexes has been motivated, in part, by the known sensitivity of the likelihood ratio chi-square statistic to large sample sizes. In this case, indexes have been developed that compare the fit of the proposed model to a baseline model of complete independence of the variables, including indexes that add a penalty function for estimating too many parameters. Examples include the non-normed fit index and the comparative fit index. Other classes of indexes have been motivated by a need to rethink the notion of testing exact fit in the population—an idea deemed by some to be unrealistic in most practical situations. The index available that tests approximate fit in the population is the root mean square error of approximation. Finally, another class of alternative indexes has been developed that focuses on the cross-validation adequacy of the model. Examples include the Akaike Information Criterion and the Expected Cross Validation Index. The use of these indexes is not without some degree of controversy (e.g., see Kaplan, 2000; Sobel & Bohrnstedt, 1985).

STATISTICAL ASSUMPTIONS UNDERLYING SEM

As with all statistical procedures, structural equation modeling rests on a set of underlying assumptions. The major assumptions associated with structural equation modeling include multivariate normality, no systematic missing data, sufficiently large sample size, and correct model specification.

Multivariate Normality

A basic assumption underlying the standard use of structural equation modeling is that observations are drawn from a continuous and multivariate normal population. This assumption is particularly important for maximum likelihood estimation because the maximum likelihood estimator is derived directly from the expression for the multivariate normal distribution. If the data follow a continuous and multivariate normal distribution, then maximum likelihood attains optimal asymptotic properties—namely, that the estimates are normal, unbiased, and efficient.

The effects of non-normality on estimates, standard errors, and tests of model fit are well known. The extant literature suggests that non-normality does not affect parameter estimates. In contrast, standard errors appear to be underestimated relative to the empirical standard deviation of the estimates. With regard to goodness-of-fit, research indicates that non-normality can lead to substantial overestimation of likelihood ratio chi-square statistic, and this overestimation appears to be related to the number of degrees of freedom of the model.

Estimators for Non-Normal Data

One of the major breakthroughs in structural equation modeling has been the development of estimation methods that are not based on the assumption of continuous multivariate normal data. Browne (1984) developed an asymptotic distribution free (ADF) estimator based on weighted least squares theory, in which the weight matrix takes on a special form. Specifically, when the data are not multivariate normal, the weight matrix can be expanded to incorporate information about the skewness and kurtosis of the data. Simulation studies have shown reasonably good performance of the ADF estimation compared to maximum likelihood (ML) and GENERALIZED LEAST SQUARES (GLS); however, ADF suffers from computational difficulties if the model is large (in terms of degrees of freedom) relative to the sample size. These constraints have limited the utility of the ADF estimator in applied settings.

Another difficulty with the ADF estimator is that social and behavioral science data are rarely continuous. Rather, we tend to encounter categorical data, and often, we find mixtures of scale types within a specific analysis. When this is the case, the assumptions

of continuity as well as multivariate normality are violated. Muthén (1984) made a major contribution to the estimation of structural equation models under these more realistic conditions by providing a unified approach for estimating models containing mixtures of measurement scales.

Missing Data

Generally, statistical procedures such as structural equation modeling assume that each unit of analysis has complete data. However, for many reasons, units may be missing values on one or more of the variables under investigation. Numerous ad hoc approaches exist for handling missing data, including LISTWISE and PAIRWISE DELETION. However, these approaches assume that the MISSING DATA are unrelated to the variables that have missing data, or any other variable in the model, that is, missing completely at random (MCAR)—a heroic assumption at best. More sophisticated approaches for dealing with missing data derive from the seminal ideas of Little and Rubin (1987).

Utilizing the Little and Rubin framework, a major breakthrough in model-based approaches to missing data in the structural equation modeling context was made simultaneously by Allison (1987) and Muthén, Kaplan, and Hollis (1987). They showed how structural equation modeling could be used when missing data are missing at random (MAR), an assumption that is more relaxed than MCAR.

The major limitation associated with the Muthén et al. (1987) approach estimator is that it is restricted to modeling under a relatively small number of distinct missing data patterns. Small numbers for distinct missing data patterns will be rare in social and behavioral science applications. Recently, however, it has been possible to incorporate MAR-based approaches (Little & Rubin, 1987) to modeling missing data within standard SEM software in a way that mitigates the problems with the Muthén et al. approach. Specifically, imputation approaches have become available for maximum likelihood estimation of structural model parameters under incomplete data assuming MAR. Simulation studies have demonstrated the effectiveness of these methods in comparison to listwise deletion and pairwise deletion under MCAR and MAR. Maximum likelihood estimation under MAR is available in LISREL, AMOS, and Mplus.

Specification Error

Structural equation models assume that the model is properly specified. SPECIFICATION error is defined to be the omission of relevant variables in any equation of the system of equations defined by the structural equation model. This includes the measurement model equations as well as the structural model equations.

The mid-1980s saw a proliferation of studies on the problem of specification error in structural equation models. On the basis of work by Kaplan and others (summarized in Kaplan, 2000), the general finding is that specification errors in the form of omitted variables can result in substantial parameter estimate bias. In the context of sampling variability, standard errors have been found to be relatively robust to small specification errors. However, tests of free parameters in the model are affected in such a way that specification error in one parameter can propagate to affect the power of the test in another parameter in the model.

MODERN DEVELOPMENTS CONSTITUTING THE "SECOND GENERATION" OF STRUCTURAL EQUATION MODELING

Recent methodological developments have allowed traditionally different approaches to statistical modeling to be specified as special cases of structural equation modeling. The more important of these modern developments have resulted from specifying the general model in a way that allows a "structural modeling" approach to other types of analytic strategies. The most recent example of this development is the use of structural equation modeling to estimate multilevel data, including models that estimate growth in longitudinal studies and the combination of latent class and latent variable modeling.

Multilevel Structural Equation Modeling

In recent years, attempts have been made to integrate MULTILEVEL modeling with structural equation modeling so as to provide a general methodology that can account for issues of measurement error and simultaneity as well as multistage sampling.

Multilevel structural equation modeling assumes that the levels of the within-group endogenous and exogenous variables vary over between-group units. Moreover, it is possible to specify a model that is assumed to hold at the between-group level, and that explains between-group variation of the within-group

variables. An overview and application of multilevel structural equation modeling can be found in Kaplan (2000).

Growth Curve Modeling From the Structural Equation Modeling Perspective

GROWTH CURVE MODELING has been advocated for many years by numerous researchers for the study of interindividual differences in change. The specification of growth models can be viewed as falling within the class of multilevel linear models (Raudenbush & Bryk, 2002), where level-1 represents intra-individual differences in initial status and growth, and level-2 models individual initial status and growth parameters as a function of interindividual differences. Recent research has also shown how growth curve models could also be incorporated into the structural equation modeling framework utilizing existing structural equation modeling software. The advantage of the structural equation modeling framework to growth curve modeling is its tremendous flexibility in specifying models of substantive interest.

Finally, recent developments by Muthén (2002) and his colleagues have led to the merging of growth curve models and latent class models for purposes of isolating unique populations defined by unique patterns of growth. This methodology constitutes *growth mixture modeling*. Extensions of growth mixture modeling allow latent class membership to be predicted by characteristics of individuals—referred to as *general growth mixture modeling*.

—David W. Kaplan

REFERENCES

- Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. C. Clogg (Ed.), *Sociological methodology 1987*. San Francisco: Jossey-Bass.
- Browne, M. W. (1984). Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62–83.
- Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York: Academic Press.
- Kaplan, D. (1990). Evaluation and modification of covariance structure models: A review and recommendation. *Multivariate Behavioral Research*, 25, 137–155.
- Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115–132.
- Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 51, 431–462.
- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29(1), 81–117.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. In C. C. Clogg (Ed.), *Sociological methodology 1987* (pp. 105–129). San Francisco: Jossey-Bass.
- Sobel, M., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), *Sociological methodology 1985* (pp. 152–178). San Francisco: Jossey-Bass.

STRUCTURAL LINGUISTICS

Structural linguistics studies language as a structure or system rather than a process. It formed the basis for the rapid development of STRUCTURALISM across the social sciences and humanities in the 1960s, and its influence can be traced through many philosophical and theoretical debates in the last decades of the 20th century.

The founder of structural linguistics was Ferdinand de Saussure. Writing around the same time as Durkheim, he treated language as a social fact—something that exists over and above the individual members of a linguistic community and imposes itself upon them. We cannot study or understand a language by looking at individual speech acts, which are unique and cannot be the object of a science. We have to look at the language that enables speech acts. Saussure argues that a language consists of a finite number of sounds and rules about the combination of these sounds, rather like the rules of grammar but working at a more fundamental level (Saussure 1974).

The basic unit of a language is the sign, made up of a *signifier*, the sound or marks on a piece of paper, and a *signified*, the concept. It is important to emphasize that the signified is the concept and not the object. The relationship between words and objects is conventional, not a matter of necessity, and we have to accept the

relationship, say, between the word “leg” and the long things at the end of (most of) our bodies if we wish to be part of our linguistic community. They could, in fact, be called anything. This breaks our common-sense relationship between words and things, and it contributed toward the “linguistic turn” in 20th-century social theory and philosophy.

The linguistic sign is combined with others into a system, and we can find the significant elements of a language by a method of *concomitant variation*, or varying each element until we find one that changes the meaning of the sign. For example, if we change the “c” in “cat” to “m,” we get a very different meaning and thus establish the opposition “c/m” as a significant one. These elements can be analyzed on two axes: the syntagmatic level, which concerns rules about what can follow what in a sentence, and the paradigmatic level, which concerns what can be substituted for what in a sentence.

This means that within the conventional system of language, each sign takes its meaning from its relationship to other signs; meaning does not exist *in* words but *between* words.

At its simplest, the structuralist movement involved the idea that the world we experience is the product of underlying structures. It was particularly useful in the analysis of cultural systems, varying from the huge project of Claude Lévi-Strauss’s (1969). *The Elementary Structures of Kinship* to William Wright’s (1975) analysis of the underlying narrative structures of different types of Western film.

—Ian Craib

REFERENCES

- Lévi-Strauss, C. (1969). *The elementary structures of kinship*. London: Eyre and Spottiswoode.
 Saussure, F. de (1974). *Course in general linguistics*. London: Fontana/Collins.
 Wright, W. (1975). *Six guns and society*. Berkeley: University of California Press.

STRUCTURALISM

Structuralism is rooted in the idea that language acquires its meaning from the juxtaposition of elements within a linguistic system. As an antecedent to the textual turn in the social sciences, structuralism lent a rigorous and objective character to the study of

culture by investigating the systemic rules and models underlying text. An emphasis on the social power of language to order individual speech and action is a hallmark of structuralist analysis. Although structuralism has been criticized for being antihumanistic and deterministic, its core insights have been incorporated into SEMIOTICS, POSTSTRUCTURALISM, POSTMODERNISM, and structuration theory.

Ferdinand de Saussure, the father of STRUCTURAL LINGUISTICS, suggested that the proper approach to language is to examine the underlying rules and models of the linguistic systems that make it possible for real-life speech to operate. Discounting the importance of both the historical evolution and the individual use of language, Saussure argued that language is a static system of collective rules (*langue*), rather than serial speech acts (*parole*). Moreover, the meaning of language is to be found in the relations among elements within the linguistic system, not in any extralinguistic considerations. These intellectual moves paved the way for the collective, systematic, rule-based study of languages and other aspects of culture.

A strong program for the autonomy of the cultural system, structuralism came to be regarded as an IDEALIST alternative to materialist perspectives such as Marxism. Anthropologists subsequently broadened the approach to encompass myth, religion, and kinship systems. Claude Lévi-Strauss found that myths could be probed for an underlying order structuring the narrative form and social life itself. Among the most significant methodological concepts is the binary opposition, a recurrent polarization of difference found in pairs of terms—such as male-female, raw-cooked, sacred-profane, purity-danger, and self-other—that Lévi-Strauss ultimately attributes to the structures of the mind itself. These oppositions impose an intelligible order on the basic data of experience through a principle of difference and exclusion. Irresolvable conflict between the binary terms motivates mythmaking, whose underlying purpose is to probe and resolve structural tensions.

A major development for structuralism was its metamorphosis into semiology, the science of signs. The key to this development was the realization that nonlinguistic systems could also be fruitfully analyzed using structuralist principles. The critic and philosopher Roland Barthes applied the principles of structuralism to a wide variety of Western cultural artifacts, notably fashion, eating, and wrestling. Barthes worked by isolating a system for study and then delineating the

semiotic codes through which the systems operate. In the case of fashion, Barthes argued that the semiotic logic of fashion constituted an autonomous “fashion system.”

Poststructuralist perspectives, underscoring contingency, action, and change within cultural systems, emerged in the 1960s partially as a criticism of structuralism’s deterministic view of culture. This acute critical reaction, whose principal contributors included Jacques Derrida (DECONSTRUCTIONISM) and Michel Foucault (genealogy), sought to understand the historical and social construction of language and signs. Poststructuralist insights into the structuring of structure were critically appropriated by Anthony Giddens and form the basis of his influential theory of structuration. Giddens makes the case that structure constrains agency while agency simultaneously gives reality to structure—a dynamic captured in the concept of practice. Structuralism’s orientation toward text and symbols also influenced postmodernism. For example, Jean Baudrillard’s explorations into the code of consumer culture owe a debt to the structuralist approach to meaning.

—George Ritzer and Todd Stillman

REFERENCES

- Barthes, R. (1967). *Elements of semiology*. London: Jonathan Cape.
- Baudrillard, J. (1998). *The consumer society*. London: Sage. (Originally published in 1970)
- Giddens, A. (1979). *Central problems in social theory*. Berkeley: University of California Press.
- Lévi-Strauss, C. (1963). *Structural anthropology*. New York: Basic Books.
- Saussure, F. de (1966). *Course in general linguistics*. New York: McGraw-Hill. (Originally published in 1915)

STRUCTURED INTERVIEW

A structured interview is an interview in which the questions the interviewers are to ask and, in many cases, the answer categories the respondents are to use have been fully developed and put in an INTERVIEW SCHEDULE before the interview begins. Structured interviews, sometimes also referred to as *standardized interviews*, are almost universally used in social SURVEYS designed to produce statistical descriptions of populations.

Structured interviews can be contrasted with SEMISTRUCTURED INTERVIEWS and UNSTRUCTURED INTERVIEWS. When an interviewer is conducting a semistructured interview, he or she has a list of topics or issues to be covered, but the specific wording of the questions is left to the interviewer. In addition, the interviewer is free to ask about additional topics or issues that were not anticipated in the interview outline. In an unstructured interview, interviewers are free during the interview to decide what topics to ask about, as well as how to word the specific questions.

The use of structured interviews is essential if the goal of the interviews is to produce statistical data. The answers to questions cannot be tabulated meaningfully unless all respondents answered the same question.

Example: Do you approve or disapprove of the way the President of the United States is doing his job?

A question similar to this is used routinely to tabulate the approval rating of presidents. Samples of people are asked this question, and the percentage of people saying they “approve” is calculated. The percent approving is tabulated over time to get a comparative sense of how people are feeling about the president, compared to other times in his administration and compared to other presidents.

Suppose, however, that some people were asked not about the president but about whether they approved or disapproved of the way the administration was conducting its affairs. Further suppose that some people were asked not whether they approved or disapproved, but whether they supported or opposed what the president was doing. These changes in wording would mean that people were generally being asked the same question, but research has shown that small changes in the wording of questions or answer categories can have a definite effect on the answers that are obtained. In short, if there was this kind of variation in the wording of questions that people answered, the tabulation of their answers would not be interpretable.

THE EVOLUTION OF STRUCTURED INTERVIEWS

In the early days of survey research, semistructured interviews were the norm. Interviewers were sent out with a list of topics, but they were free to devise their own questions to find out what people had to

say. Three considerations have led to the evolution of survey methods so that structured interviews are almost universally used:

1. Studies showed that small changes in the wording of questions or in the alternatives offered had discernible effects on the answers people gave.

2. When interviewers were free to decide on the wording of questions, not surprisingly they differed in the ways in which they worded questions to achieve the same objective. As a result, the answers obtained to surveys were often related to which interviewer did the interviewing.

3. When respondents were allowed to answer in their own words, it often was difficult to compare what they had to say. For example, if one respondent described her health as “not bad,” while another described her health as “fairly good,” could we reliably say that the first person’s health was better, worse, or about the same as that of the second person’s? Researchers have found that they can tabulate answers more reliably, and compare the answers of different respondents, if they have respondents choose from the same set of response alternatives.

Thus, in order to ensure that the differences in answers people give stem from differences in what they have to say, rather than differences in the questions they were asked or the way they chose to phrase their answers, researchers now routinely use structured interviews, in which both the wording of the questions and the response alternatives are designed ahead of time and are the same for all respondents.

STANDARDS FOR A STRUCTURED INTERVIEW

For semistructured and unstructured interviews, interviewers are free to adapt the wording of questions to the particular situation and to their perceptions of the characteristics of the respondents with whom they are talking. In a structured interview, the design of the interview schedule must create a set of questions and answers that will work “comfortably” for all respondents. This means that the design of the interview schedule must be carefully done; a structured interview schedule must meet numerous criteria in order to achieve its goals effectively:

1. The words have to be consistently understood, that is, understood in the same way by all respondents.

Questions that include unclear or ambiguous words, that can be interpreted in more than one way, will produce less good data because, in essence, some respondents will be answering different questions from others, depending on how they understand the question.

2. The questions should ask about things that all or almost all respondents can knowledgeably answer. If the interviewer is trying to collect information that people cannot recall or that they never had, the quality of the results will not be very good.

3. The wording of the questions needs to be such that interviewers can read the questions exactly as worded, without adding, subtracting, or changing words. If interviewers have to change the questions in order for them to be understood, in essence, different interviewers will be asking different questions, and the primary goal of the structured interview will be undermined.

4. Sequencing of the questions must produce a reasonable PROTOCOL for an interview interaction. The design of the instrument has to take into account the order of questions, so that respondents are not being asked the same question more than once and so that the interviewers know which questions do and do not apply to respondents.

5. When respondents are supposed to choose from a list of answers in order to answer a question, the answers need to fit the question; that is, they need to provide an exhaustive and mutually exclusive list of all the possible answers that someone might want to give.

6. Finally, if a structured interview is going to be administered in more than one language, the designers of the interview need to consider how well the questions and answers they write will translate into languages other than the original language.

INTERVIEWING PROCEDURES

A goal of structured interviews is to have the questions asked as consistently as possible for all RESPONDENTS. To accomplish this, interviewers are supposed to follow specific procedures when they conduct structured interviews:

1. They are to ask questions in the order in which they occur in the interview schedule, and they are to ask the questions exactly as they are worded.

2. When a respondent fails to answer a question completely, interviewers are to ask follow-up questions

(called nondirective PROBES) that do not increase the likelihood of one answer over others.

3. Interviewers record answers without discretion—that is, exactly as the respondent gives them.

4. Interviewers minimize personal or judgmental feedback, trying to create a professional relationship in which providing accurate answers is the priority.

TESTING A STRUCTURED INTERVIEW

Because a structured interview must meet so many standards, it is important to evaluate an interview schedule before using it in a full-scale survey. There are three kinds of evaluations that are often done:

1. Experts in the subject matter of the interview can provide input about whether or not the questions are the right ones and are covering the material that needs to be covered. Experts in survey question design often can provide information about questions that are likely to be confusing, difficult to answer, or difficult to administer in a standardized way. Experts in reading can review structured interview schedules to identify places where questions can be made clearer or simpler, or easier to understand. Bilingual reviewers can identify questions that are likely to be difficult to translate into other languages.

2. Cognitive testing of questions began to be widely done in the 1990s. Although cognitive testing takes many forms, the basic protocol is to ask respondents to answer some test questions. Then, respondents are asked to explain both how they understood the questions and how they went about answering the questions. Through this process, researchers can identify ambiguities in question wording, difficulties that respondents have in forming answers to questions, and the extent to which answers accurately reflect what respondents have to say.

3. Field pretest of structured interviews, prior to an actual survey, is the third main way that structured interview schedules are evaluated. Typically, experienced interviewers conduct structured interviews with respondents similar to those who will be included in the full-scale survey. Interviewer reports of problems with the questions constitute one way that question problems are identified. Another approach is to tape-record the pretest interviews, with respondent permission, and listen to those tapes to systematically tabulate the rates at which questions were read exactly as worded, the

rates at which respondents had trouble answering the questions, and the rates at which respondents asked for clarification.

Because interviewers are not free to change questions during interviews, it is particularly important to do a thorough job of testing structured interview schedules before they are used.

CONCLUSION

The evolution of structured interviews has been one of the important advances in survey research and the social sciences in the 20th century. The design and evaluation of structured interview schedules is among the most important elements in a well-conducted social science survey. The failure to carefully design and evaluate the questions in a structured interview before data collection can constitute one of the important sources of error and unreliability in surveys.

—Floyd J. Fowler, Jr.

REFERENCES

- Converse, J. (1987). *Survey research in the United States*. Berkeley: University of California Press.
- Fowler, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing*. Newbury Park, CA: Sage.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

STRUCTURED OBSERVATION

Structured observation (sometimes also called *systematic observation*) is a technique for data collection that has two defining characteristics. First, it is part of the broad family of *observational* techniques in which the investigator(s) gather information directly without the mediation of respondents, interviewees, and so on. Second, it is a *structured* or *systematic* technique in which data are collected according to carefully defined rules and prearranged procedures. It can be applied to a variety of aspects of behavior and interaction in a wide variety of social settings. Like any observational technique, it is necessarily selective: The researcher is attempting to focus on those elements of the situation being observed that are relevant to particular investigatory purposes.

The directness of structured observation means that a researcher can gather data that subjects are unable to provide through other procedures, such as INTERVIEWS and QUESTIONNAIRES. This may be because of the nature and detail of the information being required (how many questions a teacher asks during a lesson, how many times someone interrupts during a conversation). It may be because of limited verbal skills (e.g., of young children). This directness also relates to the way that observation can take place in natural settings that have not been created for research purposes and that are not interrupted to allow measurements to be taken.

The structured nature of these observations derives from the use of OBSERVATION SCHEDULES through which observations are recorded according to carefully predefined criteria as categories of variables that can then be used for statistical analysis. These variables and categories must be explicitly defined so that different observers can use them in identical ways, and so that the variation in accounts arising from the particular interests and perceptions of different observers is eliminated. It is in this limited but important sense that structured observation can be described as an objective approach to data collection. The focus of the study and design of the instruments are influenced by the interests and values of the researcher in the same way as any other type of study. However, the process of data collection is independent of these values and interests, and the way in which the variables have been operationalized is open and available for scrutiny. Structured observation is usually most successful as an approach when it employs low-inference measures of an essentially factual nature and involves little observer judgment and interpretation. High-inference measures, such as those involving rating scales and incorporating judgments and evaluations by the observer, are far less likely to be reliably operationalized.

LOCATING OBSERVATIONS IN TIME

A key element of structured observation is that the data incorporate a time dimension: Observations are coded in real time and can be analyzed in terms of various temporal features. There are four kinds of time-related information that investigators may wish to consider:

1. Duration: the total amount of time taken up by a category of behavior
2. Frequency: the number of individual occurrences of a category of behavior

3. Location: the position of an occurrence within the day or week or observation session
4. Sequence: the way behaviors follow on from other behaviors

The extent to which each of these kinds of information is required will determine the kind of observation system employed.

INTEROBSERVER AGREEMENT

It is essential to ensure that all observers are using the schedule in the same way, and interobserver agreement is calculated by having two observers code the same events simultaneously. The simplest measure of interobserver agreement is the Smaller/Larger (S/L) index. When two observers have coded the same observation session, the lower count of frequency or duration is divided by the higher count. The resulting value can vary between 1 and 0. It is important to note that this index shows the extent to which the observers agree on overall occurrence but not necessarily that they agree on individual instances; that is, it shows marginal, not absolute, agreement. Therefore, it does not establish the clarity and explicitness of the observation schedule, although it can show that overall counts (but not necessarily the co-occurrence of categories of different variables) are estimated consistently.

A measure of absolute agreement is the Proportion Agreement Index (PAI). This is given by dividing the number of agreements on individual observations of a pair of observers by the total numbers of observations made. It has possible values between 1.0 (total agreement) and 0.0 (no agreement). The PAI is the most commonly used observer agreement index, and satisfactory levels generally have been suggested as 0.7 or 0.8. Some researchers correct this index for the fact that observers can agree on a code by chance even if they are using the schedule in different ways. These corrections and a fuller discussion of interobserver agreement are given by Croll (1986) and Suen and Ary (1989). In the case of long-term studies, it is important that investigators reestablish levels of interobserver agreement at regular intervals to guard against observer drift.

SOME PRACTICAL CONSIDERATIONS

Structured observation is a very resource-intensive approach to data gathering. Observation procedures need extensive development and pretesting, and

observers need careful training and monitoring. An important issue is that of REACTIVITY. Most structured observation studies aim to eliminate, or at least minimize, any influence of the process of observation on what is being observed. This frequently involves preliminary periods of fieldwork, when observers are present but not collecting data to be used in the analysis, that allow the subjects of the study to acclimatize to observer presence. Assessing reactivity may also involve debriefing of both observers and (at the end of the study) research subjects, incorporating responses to the observer as a category of the observation system. Traditionally, structured observation has been a pencil-and-paper technique in which observation sheets are completed live and later prepared for computer analysis with programs such as SPSS. Recently, however, systems have been developed for coding observation using handheld computers that can then be subjected to sophisticated analysis using dedicated software (e.g., Noldus, Trienes, Hendriksen, Jansen, & Jansen, 2000).

—Paul Croll

REFERENCES

- Croll, P. (1986). *Systematic classroom observation*. Lewes, UK: Falmer.
- Noldus, L. P. J. J., Trienes, R. J. H., Hendriksen, A. H. M., Jansen, H., & Jansen, R. G. (2000). The Observer Video-Pro: New software for the collection, management, and presentation of time-structured data from videotapes and digital media files. *Behavior Research Methods, Instruments & Computers*, 32, 197–206.
- Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioural observation data*. London: Lawrence Erlbaum.

STRUCTURED, FOCUSED COMPARISON

Alexander L. George coined the expression “structured, focused comparison” to refer to a general method of theory development in qualitative, small-*N* research (George, 1979; see also George & McKeown, 1985; George & Smoke, 1974). The method is “focused because it deals selectively with only certain aspects of the historical case” (George, 1979, p. 61), and it is “structured because it employs general questions to guide the data collection and analysis in that historical case” (p. 62).

Structured, focused comparison was developed in reaction to what George saw as the shortcomings of CASE STUDY research. He argued that case study research often is not guided by clear theoretical objectives, and its idiosyncratic emphases provide a weak basis for producing cumulative knowledge. By contrast, structured, focused comparison encourages analysts to ask a set of “standardized, general questions” across cases. Regarding standardization, researchers use controlled comparisons “to assure the acquisition of comparable data from the several cases” (George, 1979, p. 62). This approach allows researchers to systematically collect comparable data for variables across two or more cases. Regarding general questions, researchers conceptualize variables at a high enough level of abstraction to be measured systematically across different cases. Only when such general concepts are employed can analysts hope to ask questions that apply to multiple cases. At the same time, however, the use of standardized, general questions does not mean that analysts cannot also inquire about the specific features of particular cases.

Structured, focused comparison was employed most explicitly in the field of international relations. George used it effectively in his work on international deterrence, showing that at least three different patterns of deterrence failure characterized international crises (George & Smoke, 1974). These patterns were discovered through a close analysis of actual cases of international crisis. George’s conclusions about the complex causes of deterrence failure challenged other theories that did not take seriously case-based evidence, including rational deterrence theory. More recently, a number of George’s students and other scholars have used these methods in their work, including Walt (1987) and Bennett (1999).

The call for structured, focused comparison paralleled a more general trend in the social sciences away from single case studies toward small-*N* research designs. Although the expression “structured, focused comparison” is used less commonly today, especially outside of international relations theory, the general concern with codifying rigorous procedures for qualitative studies focused on a small number of cases has been widely embraced. Contemporary small-*N* methodologists often describe their work as employing “systematic and contextualized comparisons,” which corresponds closely to George’s original formulation.

—James Mahoney

REFERENCES

- Bennett, A. (1999). *Condemned to repetition? The rise, fall, and reprise of Soviet-Russian military interventionism, 1973–1996*. Cambridge: MIT Press.
- George, A. L. (1979). Case studies and theory development: The method of structured, focused comparison. In P. G. Lauren (Ed.), *Diplomacy: New approaches in history, theory, and policy* (pp. 43–68). New York: Free Press.
- George, A. L., & McKeown, T. J. (1985). Case studies and theories of organizational decision making. *Advances in Information Processing in Organizations*, 2, 21–58.
- George, A. L., & Smoke, R. (1974). *Deterrence in American foreign policy: Theory and practice*. New York: Columbia University Press.
- Walt, S. M. (1987). *The origins of alliances*. Ithaca, NY: Cornell University Press.

SUBSTANTIVE SIGNIFICANCE

Substantive significance refers to whether an observed effect is large enough to be meaningful. The concept of substantive significance was developed because statistical SIGNIFICANCE TESTS can find that very small effects are significant, even they are too small to matter. Substantive significance (also called *practical significance*) instead focuses on whether an observed RELATIONSHIP, difference, or coefficient is big enough to be considered important. Substantive significance is usually gauged intuitively, with different standards in different substantive realms.

For example, one way to think about the substantive significance of a difference between two percentages is in terms of how large that difference could possibly be. The largest possible difference in recidivism rates between prisons using two different prisoner release procedures would be 100%. Finding a difference in recidivism rates of 40% or 50% or 60% would likely be considered substantively significant, whereas obtaining a difference of 2% or 3% or 4% would likely be considered unimportant even if it passes the usual statistical significance tests.

Another way to think about substantive significance is as a comparison with the maximum possible difference from a mean level. If the mean recidivism rate is 30%, for example, at most it could be lowered by these 30%. An improvement of less than 10% of that ($.10 \times .30 = 3\%$) might be considered too small to matter, regardless of whether it is statistically significant. In some fields, researchers are expected to

specify in advance how large an effect they require for substantive significance.

STATISTICAL SIGNIFICANCE tests were developed to test whether effects obtained in small SAMPLES are likely to be real. When applied to large samples, however, statistical significance tests can find that very small effects are significant. For example, a difference of .01% (e.g., 12.23% under one experimental condition versus 12.24% under another condition) might be found to be significant when the number of CASES is in the hundreds of thousands. However, such a small difference is not likely to matter in most substantive realms. Hence, the concept of substantive significance was devised to call attention to the fact that statistical significance does not suffice. Thus, the substantive significance criterion is more stringent than the notion of statistical significance, rejecting the importance of some relationships that would pass conventional significance tests.

A similar logic holds for gauging the importance of relationships. A CORRELATION of .03 would be statistically significant if it were based on several thousand observations, but it might not be large enough to be considered substantively important in many application fields.

Fowler (1985) argues that one safeguard against reporting a statistically significant but trivial effect is to test a NULL HYPOTHESIS that specifies a nonzero effect. If the null hypothesis specifies the minimum EFFECT SIZE considered substantively important, then significance testing checks whether the observed effect is significantly greater than that minimum effect size.

—Herbert F. Weisberg

REFERENCE

- Fowler, R. L. (1985). Testing for substantive significance in applied research by specifying non-zero effect null-hypotheses. *Journal of Applied Psychology*, 70, 215–218.

SUM OF SQUARED ERRORS

Many statistical methods make use of sums of squares of various kinds; one of these is the sum of squared ERRORS (SSE). Another name used for this SUM OF SQUARES is the RESIDUAL SUM OF SQUARES. The sum is often abbreviated to terms such as SSE, ESS, RSS, and SSR, where SS stands for sum of squares, E for error, and R for RESIDUAL.

The term has its origin in measurements of simple physical quantities such as lengths or weights. When such measurements were repeated, people typically did not get the same value as from the previous measurement. It was believed that the item had one fixed, true, and perhaps unknown value, and the deviation of one measurement from another was due to errors in the measurements.

Let the true value of an item be denoted μ and the i th measurement y_i . The deviation $y_i - \mu$ of the observation from the true value then becomes the error in the measurement. The magnitude of this difference may not be known, but it is still useful to think in terms of errors when measurements vary from one replication to the next.

This line of reasoning is useful when trying to answer the question of how well a statistical model, such as REGRESSION or ANALYSIS OF VARIANCE, fits the data at hand. From such a model, it is possible to compute a predicted value of the dependent variable from the observed values of the independent variables and the estimated parameter values. The difference of the observed value from the predicted value is thought to measure the error produced by the model, borrowing the “error” term from its earlier measurement meaning. For simple regression, such an error term can be expressed as

$$e_i = y_i - \hat{y}_i = y_i - a - bx_i.$$

The letter e obviously stands for error, the two y values are the observed and predicted values of the dependent variable for the i th observation, a is the estimated intercept, b is the estimated slope of the regression line, and the x value is the value of the independent variable for this individual. Similar errors can be computed for many statistical models.

Next, we want an overall sense of the magnitudes of these errors. However, the sum of the errors will equal zero, because the positive and negative errors cancel out in the sum needed for the mean. We could take the absolute value of each error, such that all the errors become positive and find that sum, and use the mean or median of these absolute values. Mathematically, absolute values are not as simple to work with as squares. The choice was made to square each error, to make each term positive, and add these squares. This gives us the so-called *error sum of squares*.

A more modern term for error is *residual*. The difference between an observed and predicted value is thought to be due to the net effect of all other variables

not included in the *sampled* data, and there is nothing “wrong” with this difference. Thus, we get the residual sum of squares instead of the error sum of squares.

—Gudmund R. Iversen

See also RESIDUAL SUM OF SQUARES

REFERENCES

- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Iversen, G. R., & Norpoth, H. (1987). *Analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-001, 2nd ed.). Newbury Park, CA: Sage.
- Lewis-Beck, M. (1980). *Applied regression: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-022). Beverly Hills, CA: Sage.

SUM OF SQUARES

Many statistical computations involve computations of sums of squares of one kind or another. Such sums are mainly used for two different purposes: to measure how different observations within a set are from each other, or as a tool in the estimation of population parameters. As the name implies, a sum of squares is a sum of squared numbers. For example, the sum of squares of the numbers 1, 2, 3, and 4 equals $1^2 + 2^2 + 3^2 + 4^2 = 30$.

When sums of squares are used to measure how different observations within a set are from each other, we usually subtract the MEAN of the observations from each observation, square the differences, and add the squares. The four observations above have a mean of 2.5; the differences become $1 - 2.5 = -1.5$, -0.5 , 0.5 , and 1.5 . When these differences are squared and added, we get the sum of squares of 5.00. This, then, is how different the four observations are from each other, as measured by the sum of squares. Such a sum of squares can be divided by its appropriate DEGREES OF FREEDOM (3) to get the VARIANCE ($s^2 = 1.67$) and thereby the STANDARD DEVIATION ($s = 1.29$) of the observations. Both of these statistics compensate for how many observations there are, and they can be compared across data sets with different numbers of observations. The reason the sum of squares for a dependent variable is not zero is because the variable has been affected by other variables.

Sums of squares are also used for the ESTIMATION of PARAMETERS, instead of the maximum likelihood method. The purpose with a sum of squares is to write the sum depending on both the data and unknown estimates of parameters. For example, the mean of the population from which the sample with observations 1, 2, 3, and 4 was taken ought to be somewhere in the middle of these observations. One way to find such a number is to ask what the value of the constant c should be in order for the sum of squares

$$(1 - c)^2 + (2 - c)^2 + (3 - c)^2 + (4 - c)^2$$

to be the smallest it can be.

For $c = 0$, we already know that the sum equals 30. But could there be some other value of c that gives a smaller sum? The trial approach, using different values of c , would be one approach. Another is to use differential calculus and find that c should be equal to the mean of the observations, here 2.5. Then, the sum of squares equals 5.00, and there is no other value of the constant that will result in a smaller sum of squares. Not surprisingly, using the sum of squares from the ORDINARY LEAST SQUARES method gives the sample mean as an estimate of the population mean.

More importantly, both analysis of variance and regression analysis make use of this least squares method to arrive at the familiar estimates for effect parameters in analysis of variance and for the coefficients in regression. For example, the formulas for the computation of slope and intercept in simple regression are found by minimizing the residual sum of squares:

$$\sum (y_i - a - bx_i)^2.$$

This calls for taking the derivatives of this mathematical function with respect to a and b , setting the resulting expressions equal to zero, and solving for a and b .

—Gudmund R. Iversen

REFERENCES

- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Iversen, G. R., & Norpoth, H. (1987). *Analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-001, 2nd ed.). Newbury Park, CA: Sage.
- Lewis-Beck, M. (1980). *Applied regression: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-022). Beverly Hills, CA: Sage.

SUMMATED RATING SCALE. See
LIKERT SCALE

SUPPRESSION EFFECT

A traditional suppression effect in a two-PREDICTOR situation, according to Horst (1941), refers to an increase in prediction of a criterion (denoted as C) by including a predictor (denoted as S) that is completely unrelated to the criterion but is related to the other predictor (denoted as P). The first predictor is labeled as a suppressor, because it suppresses (or clears out) criterion-irrelevant variance in the second predictor. Specifically, a suppressor is “a variable, which increases the predictive validity of another variable (or set of variables) by its inclusion in a REGRESSION equation. This variable is a suppressor only for those variables whose regression weights are increased” (Conger, 1974, p. 36). In general, suppression effects tend to occur when there are high correlations among predictors.

Tzelgov and Henik (1991) further refined the above definition by stating that a predictor is considered a suppressor (S) when the STANDARDIZED REGRESSION COEFFICIENT of another predictor P (β_P) is greater than its correspondent VALIDITY coefficient (r_{CP}), assuming the predictor P is scored in the direction so that it is positively related to the criterion. That is, $\beta_P > r_{CP}$, which can further be expressed as

$$\frac{(r_{CP} - r_{CS}r_{PS})}{1 - r_{PS}^2} - r_{CP} > 0,$$

or $1 - \frac{r_{PS}}{k} > 1 - r_{PS}^2$ when $k = \frac{r_{CP}}{r_{CS}}$. The latter inequality suggests that the suppression effect is determined by r_{PS} and the ratio of two validity coefficients ($\frac{r_{CP}}{r_{CS}}$).

There are three types of suppression effect discussed in the literature: classic suppression, negative suppression, and reciprocal suppression. Classic suppression is characterized by a null relationship between the suppressor and the criterion (i.e., $r_{CS} = 0$), as shown in Table 1. Negative suppression occurs when the suppressor has a negative REGRESSION COEFFICIENT (i.e., -0.07 , shown in Table 2), although it is positively related to the criterion (i.e., 0.15). Reciprocal suppression occurs when the predictor and the suppressor are

Table 1 Example of the Classic Suppression

	Validity Coefficient	Standardized Regression Coefficient
Predictor (<i>P</i>)	0.20	0.27
Suppressor (<i>S</i>)	0.00	−0.13

NOTE: $r_{PS} = 0.5, n = 400$.

Table 2 Example of the Negative Suppression

	Validity Coefficient	Standardized Regression Coefficient
Predictor (<i>P</i>)	0.40	0.43
Suppressor (<i>S</i>)	0.15	−0.07

NOTE: $r_{PS} = 0.5, n = 400$.

Table 3 Example of the Reciprocal Suppression

	Validity Coefficient	Standardized Regression Coefficient
Predictor 1	0.50	0.65
Suppressor 2	0.20	−0.43

NOTE: $r_{PS} = -0.35, n = 400$.

positively related to the criterion, but are negatively related to each other. Reciprocal suppression is characterized when regression coefficients (in their absolute values) of the suppressor and the predictor are higher than their respective relationships with the criterion (see Table 3). In contrast to the two prior suppression situations, there are no clear characteristics in the reciprocal suppression associated with the suppressor, such as the null relationship between the suppressor and the criterion or the negative regression coefficient of the suppressor. Tzelgov and Henik (1991) suggested that reciprocal suppression is better defined in terms of “the effects of these variables on the regression weights of the variables they suppress” (p. 528).

The precondition of a suppression effect, $\beta_P > r_{CP}$, discussed so far can also be applied to cases with more than three variables as well as multivariate cases such as CANONICAL CORRELATION ANALYSIS or DISCRIMINANT ANALYSIS (see Tzelgov & Henik, 1991, for a detailed discussion). Under these complex circumstances, researchers are not able to identify which predictors are

suppressor(s) and which predictors are suppressed, although the existence of a suppression effect can be verified.

Contrary to conventional wisdom, Tzelgov and Henik (1991) provided convincing evidence that suggests suppression effects occur frequently. Because criterion-irrelevant variance in a predictor (or a set of predictors) is suppressed in the regression analysis, the suppression effect improves predictability. Should the suppression effect occur in cases with more than two EXOGENOUS VARIABLES, the theoretical mechanisms of these variables should be interpreted with extreme caution. Let’s use a two-predictor example described in Table 2 to illustrate this point. We choose two predictors as the example because the suppressor can be easily identified. Assume the job performance of programmers is the criterion, and programming skills and past experience with computers are the predictor and the suppressor, respectively. Although past experience *is* positively related to job performance, the negative regression weight suggests that more experience with computers is actually related to a decrease in job performance. Given the nature of the negative suppression effect, it is clear that past experience does *not* negatively influence job performance. It is the *measure* of past experience that suppresses criterion-irrelevant variance in the measure of programming skill, artificially boosting the regression coefficient for programming skill.

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

Conger, A. J. (1974). Revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement, 34*, 35–46.

Horst, P. (1941). The role of the predictor variables which are independent of the criterion. *Social Science Research Council, 48*, 431–436.

Tzelgov, J., & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin, 109*, 524–536.

SURVEY

The social survey is a widely used method of collecting and analyzing social data for academic, government, and commercial research.

CHARACTERISTICS OF SURVEYS

Surveys are characterized by two essential elements (Marsh, 1982): the *form* of the data and the method of *analysis*.

Form of Data

Surveys produce a structured set of data that forms a variable-by-case grid. In the grid, rows typically represent CASES and columns represent VARIABLES. The cells contain information about a case's ATTRIBUTE on the relevant variable.

QUESTIONNAIRES are widely used in surveys because they ask the questions in the same way of each person and thus provide a simple and efficient way of constructing a structured data set. However, the cells in the data grid can, in principle, be filled using any number or combination of data collection methods such as interviews, observation, and data from written records (e.g., a personnel file).

Although the rows in the data grid frequently represent people, they can represent different UNITS OF ANALYSIS such as years, countries, or organizations. Where these types of units of analysis are used, the information in each column (variables) reflects information appropriate to that type of analysis unit (e.g., characteristics of a particular year, country, or organization).

Methods of Analysis

A second characteristic of surveys is the way in which data are analyzed. One function of survey analysis is to describe the characteristics of a set of cases. A variable-by-case grid in which the same type of information is collected about each case simplifies the task of description. However, survey researchers are typically also interested in EXPLANATION—in identifying causes. This is achieved by examining VARIATION in the DEPENDENT VARIABLE (presumed effect) and selecting an INDEPENDENT VARIABLE (presumed cause) that might be responsible for this variation. Analysis involves testing to see if variation in the dependent variable (e.g., income) is systematically linked to variation in the independent variable (e.g., education level). Although any such COVARIATION does not demonstrate CAUSAL RELATIONSHIPS, such covariation is a prerequisite for causal relationships.

Unlike EXPERIMENTAL research, in which variation between cases is created by the experimenter, social

surveys rely on existing variation among the cases. That is, survey analysis adopts a relatively *passive* approach to making causal inferences (Marsh, 1982, p. 6). A survey might address the question “Does premarital cohabitation increase or decrease the stability of subsequent marriage relationships?” It is not possible to adopt the experimental method in which people are RANDOMLY ASSIGNED to one of two conditions (premarital cohabitation and no premarital cohabitation) and then seeing which relationships last longer once the couples marry. The social survey adopts a passive approach and identifies couples who cohabited before marriage and those who did not and then compares their breakdown rates.

The problem in survey research is that any differences in marriage breakdown rates between the two groups could be due to differences between the two groups other than the differences in premarital cohabitation (e.g., age, religious background, ethnicity, race, marital history). Survey analysis attempts to control for other relevant differences between groups by statistically controlling or removing the influence of these variables using a range of multivariate analysis techniques (Rosenberg, 1968).

HISTORY

As a method of social research, social surveys have evolved from the primitive methods of 19th-century social reformers in Britain into a widespread and sophisticated research tool. The early surveys were conducted by British social reformers who studied the conditions of the poor affected by rapid industrialization. These surveys would be unrecognizable as social surveys today.

Modern surveys received considerable impetus from the needs for better knowledge about poverty, unemployment, and consumption patterns during the Great Depression of the 1930s and World War II. These massive social disruptions resulted in the establishment of major government surveys in Britain and the United States (e.g., the postwar Social Survey in Britain, Labour Force Surveys) and the establishment of survey research institutions such as the Bureau of the Census and National Center for Health Statistics in the United States. During World War II, the Research Branch of the U.S. Army recruited researchers such as Stouffer, Merton, Lazarsfeld, Likert, and Guttman, who became famous icons of survey research and became great innovators in the area

of survey analysis. After the war, these men established the groundbreaking university-based survey research centers that developed the methods of surveys and oversaw their increasing use in commercial and political research.

Developments in sampling methodology were crucial to the acceptance of survey research. In the early part of the 20th century, statisticians were refining sampling theory, but these developments had little impact on survey practice until the famous failures in predicting the winners of presidential elections in 1936 and 1948. The failure of NON-PROBABILITY SAMPLING led to the widespread acceptance of PROBABILITY SAMPLING methods from the 1950s onward.

The technology of social surveys has undergone substantial changes. Although structured, face-to-face interviews have been the preferred method of collecting survey information, declining response rates, expense, and safety considerations have encouraged the development of alternative data collection methods. Computerization and widespread telephone coverage have seen the widespread use of Computerized Assisted Telephone Interviews (CATI); the development of Computerized Assisted Personal Interviews (CAPI); and, more recently, Internet-based survey administration methods.

EVALUATION OF SURVEYS

Although surveys are widely accepted as a means of collecting descriptive, factual information about populations, they have faced criticisms as a means of learning about social phenomena and arriving at explanations of social processes and patterns.

Among the more important criticisms at this level are the following:

- The methods of collecting survey data are subject to considerable error stemming from sampling problems and flawed data collection instruments and methods.
- By collecting information about cases in the form of discrete variables, surveys fragment the “wholeness” of individual cases, fail to capture the meaning of attitudes and behavior for the individuals involved, and promote an atomistic way of understanding society—as a collection of discrete cases (Blumer, 1956).
- Social surveys rely on a positivist understanding of social reality, and because POSITIVISM is out of favor, surveys are condemned by association.

- The absence of experimental control and a time dimension in cross-sectional surveys means that survey research cannot adequately establish causal sequence and causal mechanisms.

Although all of these criticisms can be accurately applied to particular surveys, it is debatable whether these shortcomings are an inherent part of the survey method. Sustained research has been undertaken on ways of minimizing the sources of survey error (Groves, 1989). Marsh (1982) has argued that the criticisms in general are simply criticisms of poor surveys and argues that survey research is capable of providing an understanding of society that is adequate at the levels of both cause and meaning.

—David A. de Vaus

REFERENCES

- Blumer, H. (1956). Sociological analysis and the “variable.” *American Sociological Review*, 21, 683–690.
- de Vaus, D. (Ed.). (2002). *Social surveys* (4 vols.). London: Sage.
- Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
- Marsh, C. (1982). *The survey method: The contribution of surveys to sociological explanation*. London: Allen & Unwin.
- Rosenberg, M. (1968). *The logic of survey analysis*. New York: Basic Books.

SURVIVAL ANALYSIS

In biostatistics, survival analysis is the analysis of time-to-event (i.e., duration) data.

—Christopher Zorn

See also EVENT HISTORY ANALYSIS

SYMBOLIC INTERACTIONISM

Symbolic interactionism is a major humanistic research tradition that highlights the need for an intimate familiarity with the empirical world. Drawing upon PRAGMATISM, and having affinities with POST-MODERNISM, it shuns abstract and totalizing truths in favor of local, grounded, everyday observations. The term was coined by Herbert Blumer in 1937 and brings a major commitment to QUALITATIVE RESEARCH. It

is closely allied to George Herbert Mead's theory of the self and to the research tradition of the Chicago sociologists (especially Blumer, Park, and Hughes). Commentators have suggested a number of stages in its development, ranging from the gradual establishment of the canon between 1890 and the later 1930s on through other phases to a more recent one of diversity and new theory that has strong affinities to postmodern inquiry.

Most symbolic interactionist sociologies, their differences notwithstanding, are infused with three interweaving themes, each with methodological implications. The first suggests that what marks human beings off from all other animals is their elaborate symbol-producing capacity, which enables them to produce histories, stories, cultures, and intricate webs of communication. It is these that interactionists investigate; and because these meanings are never fixed and immutable but always shifting, emergent, and ambiguous, conventional research tools of interviews or questionnaires may not be the most appropriate way of study. Rather, an intensive engagement with the lived empirical world is called for.

This points to a second theme: that of change, flux, emergence, and process. Lives, situations, even societies are always and everywhere evolving, adjusting, becoming. This constant process makes interactionists focus on research strategies to gain access to acquiring a sense of self, of developing a biography, of adjusting to others, of organizing a sense of time, of negotiating order, of constructing civilizations. It is a very active view of the social world in which human beings are constantly going about their business, piecing together joint lines of activity and constituting society through these interactions.

This suggests a third major theme: interaction. The focus of all interactionist work is with neither the individual nor the society per se; rather, its concern is with the joint acts through which lives are organized and societies assembled. It is concerned with "collective behavior." Its most basic concept is that the self implies that the idea of "the other" is always present in a life: We can never be alone with a "self." But all of its core ideas and concepts highlight this social other, which always impinges upon the individual: The very notion of "the individual," indeed, is constructed through the other. At root, interactionism is concerned with "how people do things together" (Becker, 1986).

Unlike many other social theories, which can soar to the theoretical heavens, symbolic interactionists stay

grounded on earth. Interactionist theory can guide the study of anything and everything social, although what will be discovered is always a matter of empirical investigation. But in principle, interactionists may inspect and explore any aspect of the social world.

Some strands of interactionism (often identified as the Iowa school and linked to measurement such as the Twenty Statements Test) have tried to make the theory more rigorously operational. Others claim it to be a different kind of science (such as Anselm Strauss and the GROUNDED THEORY methodology) and seek to make it more rigorous while sustaining its distinctive commitments. Still others (often associated with Norman Denzin and postmodernism) often see the theory as needing to take a much more political and critical stand.

—Ken Plummer

See also GROUNDED THEORY, HUMANISM AND HUMANISTIC RESEARCH, INTERPRETATIVE INTERACTIONISM, INTERPRETIVISM, POSTMODERN ETHNOGRAPHY, PRAGMATISM, QUALITATIVE RESEARCH, TWENTY STATEMENTS TEST

REFERENCES

- Becker, H. S. (1986). *Doing things together*. Evanston, IL: Northwestern University Press.
- Blumer, H. (1969). *Symbolic interactionism: Perspective and method*. Englewood Cliffs, NJ: Prentice Hall.
- Denzin, N. (1992). *Symbolic interactionism: The politics of interpretation*. Oxford, UK: Blackwell.
- Gubrium, J. F., & Holstein, J. A. (1997). *The new language of qualitative method*. New York: Oxford University Press.

SYMMETRIC MEASURES

Symmetric measures of ASSOCIATION describe the relationship between two variables *X* and *Y* without differentiating if either variable is an antecedent (or independent variable) or a consequent (or dependent variable). Examples of symmetric measures include PEARSON CORRELATION COEFFICIENT, SPEARMAN CORRELATION, Goodman and Kruskal's GAMMA COEFFICIENT, and point biserial correlation coefficient, in addition to the five measures to be presented below.

Table 1 Distributions of Gender (X) and Successfulness Records (Y)*Exhibit A*

	Failure ($Y = 0$)	Success ($Y = 1$)	Totals
Female ($X = 1$)	$20 = a$	$5 = b$	$25 = (a + b)$
Male ($X = 0$)	$10 = c$	$15 = d$	$25 = (c + d)$
Totals	$30 = (a + c)$	$20 = (b + d)$	$50 = a + b + c + d$

$$\phi = \frac{bc - ad}{\sqrt{(a + c)(b + d)(a + b)(c + d)}} = \frac{5 \times 10 - 20 \times 15}{\sqrt{30 \times 20 \times 25 \times 25}} = -.41$$

Exhibit B

X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
1	1	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0
1	1	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0
1	1	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0
1	1	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0
1	1	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0

$$\sum XY = 5, \sum X = 25, \sum X^2 = 25, \sum Y = 20, \sum Y^2 = 20$$

$$\phi = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right) \left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}} = \frac{5 - \frac{25 \times 20}{50}}{\sqrt{25 - \frac{25^2}{50}} \sqrt{20 - \frac{20^2}{50}}} = -.41$$

NOTE: $X = 1$ (female), $X = 0$ (male); $Y = 1$ (success), $Y = 0$ (failure).

PHI COEFFICIENT

Phi coefficient (ϕ) is a special case of the Pearson correlation coefficient. When both variables, X and Y , are naturally dichotomous (e.g., male/female, success/failure, survival/death), the Pearson correlation is conventionally labeled as the phi coefficient. Because of its direct relationship with the Pearson correlation coefficient, the null hypothesis test and computation formula for the Pearson correlation coefficient can be applied to the phi coefficient.

Given the dichotomous nature of both variables (e.g., gender and successfulness), the distribution of gender and successfulness can be arranged in a table with four cells, often referred to as a CONTINGENCY TABLE, as shown in Exhibit A of Table 1. As seen, there are 20 successful females, 5 females who fail, 10 successful males, and 15 males who fail. Based on a simplified Pearson correlation formula,

$$\phi = \frac{bc - ad}{\sqrt{(a + c)(b + d)(a + b)(c + d)}}$$

[assuming $(a + c)(b + d)(a + b)(c + d) \neq 0$], the correlation between gender (X) and successfulness (Y)

is $-.41$, if females are CODED as 1 and males are coded as 0 for variable X , and success is coded as 1 and failure is coded as 0 for variable Y . Alphabetical letters a , b , c , and d represent the frequencies in each cell of the contingency table. We can also calculate the phi coefficient based on the conventional Pearson correlation formula. As demonstrated in Exhibit B of Table 1, the correlation between gender and successfulness is $-.41$. This negative correlation suggests that the higher the “gender” (from 0 to 1), the lower the “successfulness” (from 1 to 0). In other words, fewer females succeed than do males.

TETRACHORIC CORRELATION

If the phi coefficient is calculated based on two artificially dichotomized variables, actually CONTINUOUS VARIABLES with bivariate NORMAL DISTRIBUTIONS, the phi coefficient will underestimate the relationship between the variables. To estimate what the actual relationship would be if both variables were not artificially dichotomized, we could apply the tetrachoric correlation formula. For example, two variables, sleep (X) and

Table 2 Distributions of Sleep (X) and School Performance (Y)

	Poor Performance ($Y = 0$)	Good Performance ($Y = 1$)	Totals
Long Sleep ($X = 1$)	60 = a	40 = b	100 = ($a + b$)
Short Sleep ($X = 0$)	40 = c	60 = d	100 = ($c + d$)
Totals	100 = ($a + c$)	100 = ($b + d$)	200 = $a + b + c + d$

school performance (Y), are artificially dichotomized so that sleep is classified as either short sleep (if a person sleeps less than 7 hours) or long sleep (if a person sleeps equal to or more than 7 hours), and school performance is classified as either poor (if a grade point average is lower than 2.5) or good (if a grade point average is equal to or greater than 2.5). The frequencies for these two variables are summarized in the contingency table depicted in Table 2.

The tetrachoric correlation formula is defined as

$$r_{\text{tet}} = \frac{bc - ad}{\lambda_X \lambda_Y n^2},$$

where n is the total SAMPLE size, λ_X is the ordinate of the standardized normal distribution at

$$\frac{a + b}{a + b + c + d}$$

(i.e., proportion of subjects obtaining 1 for variable X), and λ_Y is the ordinate of the standardized normal distribution at

$$\frac{b + d}{a + b + c + d}$$

(i.e., proportion of subjects obtaining 1 for variable Y). Based on the data in Table 2, the tetrachoric correlation is computed as

$$r_{\text{tet}} = \frac{40 \times 40 - 60 \times 60}{.398942 \times .398942 \times 200^2} = -.31.$$

When compared to r_{tet} , ϕ based on the same data is much smaller ($\phi = -.20$). It should be noted that r_{tet} is an appropriate estimate of a relationship only for sample sizes greater than 400.

In the next two sections, we will describe two related measures of association, the contingency coefficient and Cramér's V . Both measures do not require any assumptions pertaining to the nature of the variables (e.g., normal vs. SKEWED, continuous vs. DISCRETE, ordered vs. unordered). Despite this advantage, any

pair of contingency coefficients or Cramér's V coefficients are not comparable unless the values of $\min(r, c)$ in both contingency tables are the same, where $\min(r, c)$ refers to the smallest number of rows (r) or columns (c) in a contingency table.

CONTINGENCY COEFFICIENT

We often conduct the CHI-SQUARE TEST (χ^2) to examine if there is an association between two categorical variables, which are presented in the form of a contingency table. For instance, a car dealer examines whether there is an association between color preference (e.g., white, black, gold, blue) and buyer ethnicity (e.g., Caucasian, Hispanic, African American) according to her customer database. Although she can apply the χ^2 test to see if any association between these two variables exists, the χ^2 value does not provide information about the magnitude of the relationship because two different χ^2 values may suggest the same magnitude of association. To address this limitation, the contingency coefficient, C , has long been applied, where

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}},$$

and n is the total observed frequencies in the contingency table. By carefully examining the formula, it is clear that the contingency coefficient cannot approach unity even if there is a perfect association. Assuming there is a perfect association, the coefficient is equal to

$$\sqrt{\frac{c - 1}{c}}$$

if the contingency table has the same number of columns (c) and rows (r).

CRAMÉR'S V

Cramér's V was developed to resolve the limitation that the contingency coefficient cannot reach unity.

Conceptually, Cramér's V coefficient,

$$V = \sqrt{\frac{\chi^2}{n \times [\min(r, c) - 1]}}$$

is viewed as a ratio of an observed statistic to its maximum statistic. Therefore, the coefficient can range from 0 to 1. The null hypothesis test pertaining to both Cramér's V coefficient and the contingency coefficient can be conducted based on the χ^2 test of independence, with the DEGREES OF FREEDOM equal to $(r - 1)(c - 1)$.

KENDALL'S TAU

A rival of the Spearman correlation coefficient, Kendall's tau (τ) coefficient assesses the relationship between two ordinal variables, X and Y , based on n cases. Conceptually, it assesses the proportion of discrepancy between numbers of concordance and discordance expressed by

$$\tau = \frac{P - Q}{\frac{n(n-1)}{2}}.$$

P is the number of concordance, and Q is the number of discordance. (X_i, Y_i) from case i and (X_j, Y_j) from case j are considered to be concordant if $Y_i < Y_j$ when $X_i < X_j$, or if $Y_i > Y_j$ when $X_i > X_j$, or if $(X_i - X_j)(Y_i - Y_j) > 0$. Similarly, (X_i, Y_i) and (X_j, Y_j) are viewed to be discordant if $Y_i < Y_j$ when $X_i > X_j$, or if $Y_i > Y_j$ when $X_i < X_j$, or if $(X_i - X_j)(Y_i - Y_j) < 0$. Kendall's τ coefficient can range from -1 when $P = 0$ to 1 when $Q = 0$. Where there are tied ranks within variable X or variable Y , the maximum absolute value of Kendall's τ coefficient will be less than 1. Under this circumstance, researchers can employ either Kendall's τ_b coefficient or Stuart's τ_c coefficient. Both indexes differ in terms of how tied ranks are treated. Examples of applying these three coefficients are presented in Chen and Popovich (2002).

The NULL HYPOTHESIS test of $\tau = 0$ can be conducted by a z -test,

$$z = \frac{\tau}{\frac{\sqrt{2(2n+5)}}{3\sqrt{n(n-1)}}},$$

if the sample size is greater than 10. In contrast, a t -test shall be performed to examine the null hypothesis for both τ_b and τ_c (cf. Kendall & Gibbons, 1990).

—Peter Y. Chen and Autumn D. Krauss

REFERENCES

- Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage.
- Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.
- Kendall, M., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). New York: Oxford University Press.
- Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

SYMMETRY

Symmetry is a general property referring to the shape of a DISTRIBUTION, or to a COEFFICIENT that makes no distinction between the INDEPENDENT and DEPENDENT VARIABLE. A NORMAL DISTRIBUTION is symmetric, being bell-shaped with an equal number of cases scoring above and below the mean. A symmetric distribution is not SKEWED to the right or left. A coefficient is symmetric if, in its calculation, it is not necessary to declare variables independent or dependent. The PEARSON CORRELATION COEFFICIENT is symmetric, in that when estimated for two variables, Q and Z , the resulting value will be exactly the same value, regardless of whether Q is considered independent and Z dependent, or vice versa. In contrast, the coefficient LAMBDA is ASYMMETRIC, in that its value changes depending on which of the two variables is considered independent. Finally, it should be noted that symmetry is sometimes used to refer to the shape of a social structure, as in "network symmetry."

—Michael S. Lewis-Beck

SYNCHRONIC

Synchronic features describe phenomena at one point in time, implying the researcher employs a cross-sectional, snapshot, or static perspective (cf. DIACHRONIC).

—Tim Futing Liao

SYSTEMATIC ERROR

Systematic error is the consistent over- or underestimation of a PARAMETER. When the effects of systematic

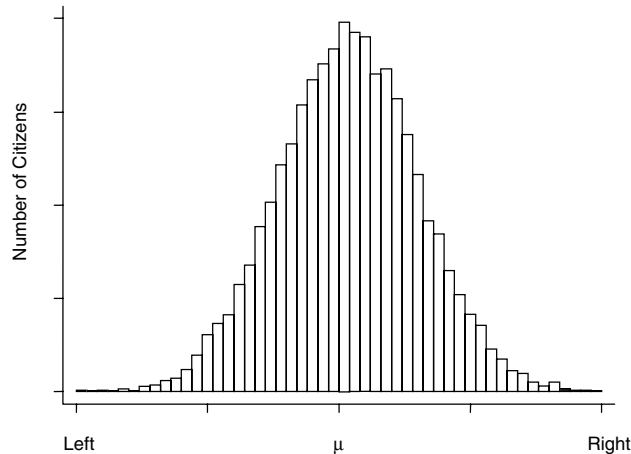


Figure 1 Random Error

error are unaccounted for or ignored, analysts will draw incorrect conclusions from their research.

Analysts distinguish systematic errors from RANDOM ERRORS in the following way (see Walker & Lev, 1953): Suppose multiple measurements of a characteristic with a true value μ are taken. Some measurements generate estimates that are too large, whereas others are too small. If the instrument is unbiased, however, the positive and negative errors will cancel each other out, and estimates will be correct on average, centered at μ (see Figure 1).

Alternatively, a faulty instrument, for example, produces systematic error by reporting measurements that are, on average, larger (or smaller) than μ (see Figure 2).

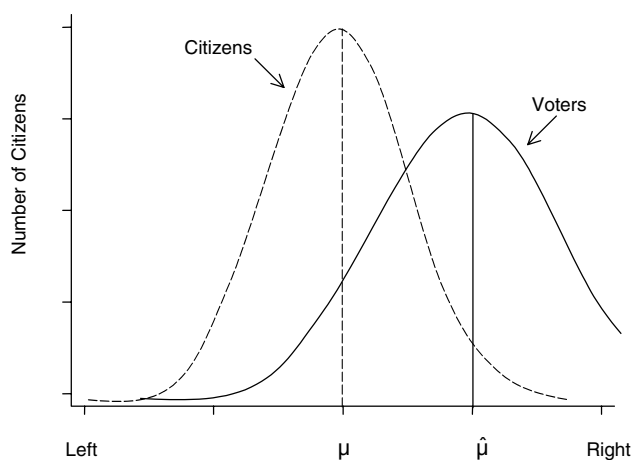


Figure 2 Systematic Error

Repeated measurements are not centered at μ , but at the overestimated $\hat{\mu}$. Analysts should note that systematic error may affect all (as seen in Figure 2) or some proportion of measurements taken, or may affect different subsets of observations to different extents (Nunnally & Bernstein, 1994). In each of these cases of systematic error, a researcher's measurements are misleading: Whereas random error affects the precision of the estimate, systematic error biases the parameter estimate itself.

To illustrate, consider the following example: In democracies, a population's party preferences are measured through elections. Figure 1 reports the distribution of preferences in a two-party system: The horizontal axis indicates the strength of identification with a particular party ("Right"), and the vertical axis reports the number of citizens at a particular identification level. μ reports the true proportion of citizens who prefer the other party, "Left" to Right. However, elections are imperfect instruments for the measurement of a population's preferences. Suppose that voters who prefer Left are less likely to vote than those who prefer Right. In this case, the distribution of voters resembles that depicted by the solid line in Figure 2, which is centered at $\hat{\mu}$ (see Franzese, 2002) rather than the dotted line centered at μ . Conclusions about the distribution of preferences in a population based on this election result are likely to overestimate the popularity of Right.

This example suggests one type of systematic error, SELECTION BIAS. The selection of observations (voters rather than citizens) results in the miscalculation of support for Right within the population. However, several other types of systematic error may occur in social science research. Measurement error results when faulty instruments are used; the randomization of the order of response categories in survey research, for example, is an attempt to limit measurement error in survey response, because participants' answers have been shown to depend on the order in which response categories are presented (e.g., Wiseman, Moriarty, & Schafer, 1975–1976). CONFOUNDING error, another type of systematic error, occurs when an estimate of the relationship between two variables of interest is biased by the exclusion of a third, perhaps unobserved variable (or "confounder"). If, for example, Right party supporters are, on average, older than those who identify with the Left Party, there may appear to be a relationship between age and support for the

government policy. However, when the party identification is considered (see Figure 3), it becomes clear that party affiliation, regardless of age, is associated with attitudes toward the government policy.

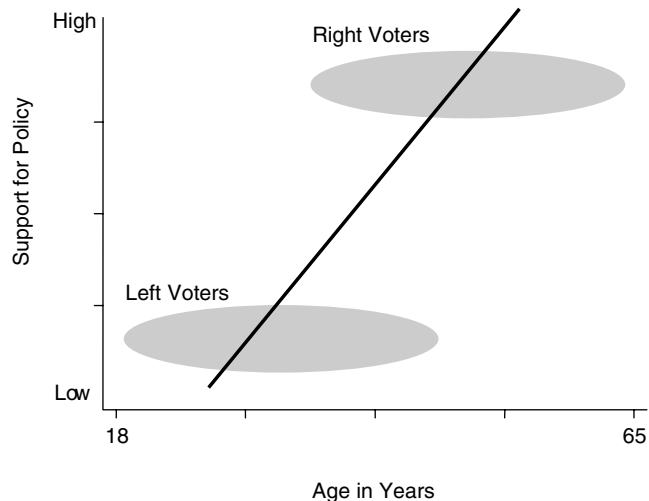


Figure 3 Types of Systematic Error: Confounding Error

Therefore, analysts are well advised to take these and other sources of systematic error into account when constructing their models, or to anticipate the effects of systematic error when drawing their conclusions.

—Karen J. Long

REFERENCES

- Franzese, R. (2002). *Macroeconomic policy of developed democracies*. Cambridge, UK: Cambridge University Press.
- Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Walker, E., & Lev, J. (1953). *Statistical inference*. New York: Henry Holt.
- Wiseman, F., Moriarty, M., & Schafer, M. (1975–1976). Estimating public opinion with the randomized response model. *Public Opinion Quarterly*, 39(4), 507–513.

SYSTEMATIC OBSERVATION. See STRUCTURED OBSERVATION

SYSTEMATIC REVIEW

Systematic review describes a specific methodology for conducting reviews of literature. This methodology prescribes explicit, reproducible, and transparent processes for collating the best available evidence in answer to specific questions. In particular, it requires the use of robust techniques for searching for and identifying primary studies, appraising the quality of these studies, selecting the studies to be included in the review, extracting the data from the studies, and synthesizing the findings narratively and/or through pooling suitable quantitative data in META-ANALYSIS. Systematic review has developed in response to the perceived problems of traditional “narrative” reviews, where haphazard and idiosyncratic techniques, such as nonsystematic methods of literature identification and informal summary, can introduce bias. In health care, these problems had led to delayed introduction of effective treatments and continuation of ineffective treatments. Increasingly, systematic review extends beyond evaluation of interventions to embrace a diversity of RESEARCH QUESTIONS.

Systematic review has benefited from the development of electronic databases, allowing for more extensive and precise searching for candidate articles. Importantly, systematic reviews may, to reduce the risk of publication biases, also seek unpublished studies.

Systematic reviews critically appraise the quality of the relevant primary research, sometimes facilitated by checklists and scoring systems. RANDOMIZED CONTROL TRIALS (RCTs) are often seen as the “gold standard” of study design for evaluating interventions and are frequently found at the top of “hierarchies of evidence,” which grade study designs according to the RELIABILITY of the evidence they are likely to produce. However, conducting RCTs may not always be possible for practical, ethical, or economic reasons; sometimes, it is more appropriate or necessary to use non-RCT evidence.

Once the studies to be included in the review have been selected, and relevant data extracted, some form of synthesis can be attempted. Meta-analysis refers to statistical methods for pooling quantitative results from different studies (Egger, Davey Smith, & Altman, 2001). Sometimes, meta-analysis is used as a synonym for systematic review, but generally, a meta-analysis will form only part of a systematic review, and many

systematic reviews do not include meta-analyses. The feasibility of extending meta-analysis to include qualitative evidence using Bayesian statistical methods has recently been investigated. More usually, qualitative evidence, and quantitative evidence not meta-analyzed, are summarized narratively.

Since its introduction in the 1980s, systematic review has become central to the evidence-based practice movement in the health sciences. It is the core activity of the Cochrane Collaboration, an international network of groups producing systematic reviews of health interventions. The more recently established Campbell Collaboration will conduct systematic reviews in areas of social, behavioral, educational, and criminological intervention.

Well-conducted systematic reviews have already yielded significant benefits. The methodology has had only a short time to develop, and it will gain from further development and refinement. The frequent exclusion from some systematic reviews of some relevant data sources (e.g., from observational and qualitative studies) has been criticized (Dixon-Woods, Fitzpatrick, & Roberts, 2001). Recent guidance on the conduct of systematic reviews (NHS Centre for Reviews and Dissemination, 2001) has recognized these problems, and methods are now rapidly evolving to extend the science of systematic review to incorporate more diverse forms of evidence.

—Mary Dixon-Woods and Alex Sutton

REFERENCES

- Dixon-Woods, M., Fitzpatrick, R., & Roberts, K. (2001). Including qualitative research in systematic reviews: Problems and opportunities. *Journal of Evaluation in Clinical Practice*, 7, 125–133.
- Egger, M., Davey Smith, G., & Altman, D. G. (2001). *Systematic reviews in health care: Meta-analysis in context* (2nd ed.). London: BMJ Books.
- NHS Centre for Reviews and Dissemination. (2001). *Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews* (Report number 4, 2nd ed.). York, UK: CRD.

SYSTEMATIC SAMPLING

Systematic sampling involves selecting units at a fixed interval. There are two common applications. The

first is where there exists a list of units in the population of interest that can be used as a SAMPLING FRAME. Then, the procedure is to select every A th on the list. The second is where no list exists, but sampling of a flow is to be carried out in the field. (Examples of flows include visitors to a museum, passengers arriving at a station, and vehicles crossing a border.) Then, the procedure is to select every A th unit passing a specified point. This is conceptually equivalent to creating a list of units in the order that they pass the point and then SAMPLING systematically from the list, so the principles of systematic sampling can be illustrated by considering only list sampling.

The first step in systematic sampling is to calculate the sampling interval, $A = \frac{N}{n}$. For example, to select a sample of $n = 100$ from a list of $N = 1,512$ units, then $A = 15.12$. The second step is to generate a random start. This will determine which is the first unit on the list to be sampled. In fact, it will determine the entire sample, because the remaining selections are obtained by repeatedly adding the fixed interval; no further random mechanism is involved. The random start should be a number between 1 and $1 + A$. Step 3 is to successively add A to the random start. Thus, if the random start is 4.86, the following numbers will be obtained: 4.86, 19.98, 35.10, 50.22, . . . , 1501.74. The final step is to ignore the decimal places—the sampled units are the 4th, 19th, 35th, 50th, . . . 1501st on the list. This procedure will result in the selection of exactly n units. In the case of flow sampling, N is unknown, so A is usually chosen on practical grounds.

A common motivation for using systematic sampling is practical simplicity. The case of flow sampling is a good example. In many situations, there is no other way to select a random sample from a flow, because no lists exist and it is not possible to create such a list in order to sample subsequently from the list in some other way. Practical considerations also apply in other situations. For example, consider sampling from a set of card files arranged in a series of filing cabinets. Systematic sampling avoids the need to remove the cards from the cabinets and re-sort them, or to number all the cards. It is necessary only to count through them.

If the units are ordered in a meaningful way, systematic sampling has the (typically beneficial) effect of achieving implicit stratification (see STRATIFIED SAMPLE). This is often perceived as an advantage of systematic sampling over SIMPLE RANDOM SAMPLING.

Many surveys employ systematic sampling within explicit strata. Again, the same argument applies to systematic flow sampling. If a survey is sampling visitors to a museum and there is a tendency for visitors with different characteristics to be represented in different proportions at different times of day, then systematic sampling over the course of a whole day should result in improved precision of estimation compared with a simple random sample of the same size of that day's visitors.

—Peter Lynn

REFERENCES

- Levy, P. S., & Lemeshow, S. (1991). *Sampling of populations: Methods and applications*. New York: Wiley.
- Scheaffer, R. L., Mendenhall, W., & Ott, L. (1990). *Elementary survey sampling*. Boston: PWS-Kent.

T

TAU. See SYMMETRIC MEASURES

TAXONOMY

Taxonomy is a term that is particularly associated with biology and refers to classification. Taxonomies (e.g., types of society, political systems) are often controversial in terms of issues such as whether the classification categories are innate or fabricated by the taxonomist.

—Alan Bryman

TCHEBECHEV'S INEQUALITY

Sometimes spelled Chebyshev's Inequality, the inequality allows for the derivation of probability bounds if only the MEAN and VARIANCE of a RANDOM VARIABLE are known. Its primary importance is in terms of its use as a tool for deriving more theoretical results (e.g., the weak LAW OF LARGE NUMBERS).

Mathematically, for random variable X with finite mean μ and variance σ^2 , then for any $y > 0$, the (two-sided) Tchebechev's Inequality is given by

$$P\{|X - \mu| \geq y\} \leq \frac{\sigma^2}{y^2}.$$

Intuitively, Tchebechev's Inequality indicates that if the variance σ^2 of a random variable is small, then X will be very close to μ .

To illustrate how Tchebechev's Inequality can be used to compute population bounds when the distribution of the random variable is unknown, consider a hypothetical case. If the average grade in a class is 75 with a variance of 20, the upper bound on the probability that a randomly selected student will have a grade below 65 or above 85 is

$$P\{|X - 75| \geq 10\} \leq \frac{20}{100} = 0.20.$$

Because no information about the distribution is used to calculate the bound, the upper bound may not be particularly accurate. For example, if we knew that the distribution of class grades was the NORMAL DISTRIBUTION, the actual probability of the above event is approximately 0.026—quite far from the bound of 0.20 provided by Tchebechev's Inequality.

Defining $y = t\sigma$ permits Tchebechev's Inequality to bound the probability that X is no more than t standard deviations away from its mean, μ :

$$P\{|X - \mu| \geq t\sigma\} \leq \frac{1}{t^2}.$$

Continuing the above example, the upper bound on the probability that the grade of a randomly selected student will be more than two standard deviations away from the class mean is less than or equal to 0.25.

It is also possible to derive the one-sided inequality for any $y > 0$:

$$P\{X \geq y\} \leq \frac{\sigma^2}{\sigma^2 + y^2}.$$

Although the ability to compute probability bounds for a random variable without knowing its distribution

is useful, Tchebechev's Inequality is most important as a tool for deriving a weak law of large numbers and for proving that convergence in mean square implies convergence in probability.

—Joshua D. Clinton

REFERENCES

- Amemiya, T. (1994). *Introduction to statistics and econometrics*. Cambridge, MA: Harvard University Press.
- Rice, J. A. (1995). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury.
- Ross, S. (1997). *A first course in probability* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

T-DISTRIBUTION

The *t*-distribution is the DISTRIBUTION for the *t*-variable. The British statistician and brewmaster William Gosset introduced the *t*-variable around 1900. He found the distribution by what today we would call SIMULATION. He made a histogram of 750 quantities he had computed and observed that the HISTOGRAM deviated some from a NORMAL DISTRIBUTION. He called the new distribution the *t*-distribution and the variable the *t*-variable. Later, Sir Ronald A. Fisher derived the mathematical formula for the distribution. A *t*-test uses the statistical *t*-variable to find the *P* VALUE for a statistical test.

The *t*-variable is a family of variables, each with its own distribution. All of the distributions are symmetric and bell shaped, but they are more spread out than the normal distribution. The variables and their distributions are numbered 1, 2, 3, and so on. This number is known as the DEGREES OF FREEDOM for the variable. The larger the degrees of freedom, the more the distribution resembles the normal distribution.

A *t*-test can be used for many different null hypotheses. The most common *t*-tests are for a single MEAN, the difference between two means, and a REGRESSION COEFFICIENT. For all these cases, the test statistic is a linear combination of the observations on the DEPENDENT VARIABLE.

t-test for a single population mean. The NULL HYPOTHESIS states $H_0: \mu = \mu_0$. The population mean is denoted μ , and the hypothesis states that the

population mean is equal to a numerical value μ_0 . The test statistic *t* is computed according to the following formula:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}.$$

Here, $\bar{y}$ is the SAMPLE mean of the variable, *s* is the sample STANDARD DEVIATION of the variable, and *n* is the number of observations in the sample. The value of the *t*-variable computed this way is for the *t*-variable with *n* – 1 degrees of freedom. Using *t* and the degrees of freedom, we can then find the corresponding *P* VALUE from tables of the *t*-distribution or from statistical software.

When the sample is small, say, less than 30 observations, the formula for the test statistic works best when the variable *Y* follows a normal distribution. For large samples, this requirement is not necessary.

t-test for the difference between two population means μ_1 and μ_2 . The null hypothesis states $H_0: \mu_1 - \mu_2 = 0$ (or any other numerical value).

The corresponding *t*-value has $n_1 + n_2 - 2$ degrees of freedom. The test works for any numerical value of the difference between the two means, not just 0. The formula for *t* can be found in any statistics textbook, and it is built into almost all statistical computer packages.

The test works best when the variable follows a normal distribution in each population and when the VARIANCES in the two distributions are equal. When the sample sizes are larger than, say, 30, these requirements are less important.

t-test for a regression coefficient β . In a regression study of an independent variable *X* and a dependent variable *Y*, we can test $H_0: \beta = 0$. The corresponding *t*-value has *n* – 2 degrees of freedom. The test works best when the *Y* values are distributed normally around the regression line.

—Gudmund R. Iversen

REFERENCES

- Iversen, G., & Gergen, M. (1997). *Statistics: The conceptual approach*. New York: Springer-Verlag.
- Moore, D. S. (1997). *Statistics: Concepts and controversies*. New York: W. H. Freeman.
- Utts, J. M. (1996). *Seeing through statistics*. Belmont, CA: Duxbury.

TEAM RESEARCH

Team research refers generally to research in which two or more people collaborate sufficiently to share authorship of findings disseminated to others, most frequently through scholarly publication. It may involve one or a stream of projects. It is distinguished from single-author research, including research authored by one person but conducted with the assistance of others, in that responsibility for a complete research project, from initial conceptualization through dissemination of findings, is shared to some extent by those collaborating on the research.

Team research may take a wide variety of forms. The most standard involves two or more scholars working together on a research project and publishing one or more discrete papers on it. The scholars may be from the same institution and academic discipline or different institutions and disciplines; they may be faculty, students, or scholar-practitioners; they may be in the same location or multiple locations; and they may have similar or differing conceptual and methodological strengths.

A more complex form includes two or more scholars working together as a team on a stream of research over an extended period of time. They may view themselves as a team even if some of the individual studies of the research stream are published individually or with others who are not part of the team.

Another more complex form involves practitioner members of a setting sharing responsibility with one or more scholars for studying the setting and disseminating the research findings both externally and internally. This may take the form of ACTION RESEARCH, in which addressing a particular practical concern in the setting and scholarly dissemination of findings from addressing it are both aims.

There are both product and social motivations for conducting team research. Product motivations include the belief that the contributions of multiple authors, often with different types of knowledge, are necessary to address particular research questions adequately and creatively. Social motivations include desires to work with friends and to help develop younger scholars by including them in research. A motivation for action research is the ability to make both practical and scholarly contributions related to a particular problem in a setting.

HISTORICAL DEVELOPMENT

Research collaboration between scholars, particularly in the hard sciences and medicine, appears to have developed initially in France toward the end of the 18th century. Such collaboration grew substantially after World War I. Team research in the social sciences, in both its scholarly and action research forms, has been recognized as a distinguishable form of research at least since the 1940s, although it is less frequent than team research in the hard sciences (Bayer & Smart, 1991).

CHARACTERISTICS AND SKILLS NEEDED

Team research evokes all the dynamics of any group, including developing roles, leadership, norms, cohesion, and a balance between task and social dimensions (Del Monte, 2000). For team research to succeed, several skills and activities on the part of team members are required. In particular, this type of research involves considerable social competence, especially in dealing with potential conflicts and negotiating agreements.

There are several aspects about which it is crucial for team members to reach agreement. These include the hoped-for outcomes of the research, how the work and responsibility for the individual study and/or stream of research will be divided, the research questions to be explored, how to gather and analyze the data, who owns the data that are gathered, how to interpret the meanings of the results, confidentiality and decisions regarding sharing information with research participants, the means by which outcomes of the research will be disseminated, and the order of authorship of publications that might result from the research. In action research, it is also necessary to reach agreement regarding the means by which both practical and scholarly purposes are to be pursued.

Reaching agreement on dimensions like these is particularly important and difficult when the research team includes members who operate out of different cultural values (e.g., practice vs. scholarship, qualitative vs. quantitative methodological emphases); differ in power, status, and knowledge about how to do the type of work that will be involved; and/or differ on which aims are primary for them. Moreover, communication regarding these issues and updated agreements will likely need to occur on an ongoing basis throughout the research endeavor. This is especially the case for research teams whose expectations of joint research extend beyond an individual project.

POTENTIAL BENEFITS

There are several potential benefits of team research. One is the advantage of multiple and differing viewpoints, which, other things being equal, typically lead to a more creative approach to the research. Another benefit is that researchers can address larger problems than would be the case if they were working independently. Yet another is that team approaches to research often result in improved studies. For example, in many fields, the more prestigious journals typically have a higher proportion of collaboratively authored papers than do other publications (Bayer & Smart, 1991). There also tends to be a positive relationship between coauthorship (as opposed to single authorship) and both acceptance and citation of articles. Furthermore, even though team research may be difficult, it may also be challenging and exciting, at least when its social dynamics are handled well.

POTENTIAL PROBLEMS

Although there are multiple potential benefits, there are also many potential problems that arise in team research. It is often necessary to resolve disagreements regarding the order of authorship or inclusion of authors, and, more generally, how to allocate credit fairly among different team members. Sometimes, there are concerns about social loafing on the part of some members of the research team who do not carry through on their responsibilities. In addition, especially when multiple disciplines or multiple cultures are present within the team, differences in values, experiences, norms, and communication styles among team members may all cause conflicts, some of which may not be resolvable in a collaborative way. All of these potential problems are exacerbated when there are power and status differentials among members of the research team. If these problems are not addressed satisfactorily, the expected benefits of team research will not be experienced.

EXAMPLES

Examples of team research in which two or more scholars work together are ubiquitous in virtually all social science fields. Kirkman, Rosen, Gibson, Tesluk, and McPherson (2003), along with others, illustrate a team carrying out an extended stream of research—in their case, on virtual teams. Bartunek and Louis (1996) present many examples of insider members in education, community settings, and work

organizations collaborating in research on the setting with external researchers. Such collaboration may include an action research component, but this is not necessary.

—Jean M. Bartunek

REFERENCES

- Bartunek, J. M., & Louis, M. R. (1996). *Insider-outsider team research*. Thousand Oaks, CA: Sage.
- Bayer, A. E., & Smart, J. C. (1991). Career publication patterns and collaborative “styles” in American academic science. *Journal of Higher Education*, 62, 613–636.
- Del Monte, K. (2000). Partners in inquiry: Ethical challenges in team research. *International Social Science Review*, 75, 3–14.
- Kirkman, B. L., Rosen, B., Gibson, C. B., Tesluk, P. E., & McPherson, S. O. (2003). Five challenges to virtual team success: Lessons from Sabre, Inc. *Academy of Management Executives*, 16(3), 67–79.

TELEPHONE SURVEY

Surveys by telephone continue to be the most frequently utilized technique of asking a standardized set of questions with the purpose of gathering information from POPULATIONS, large and small, for a variety of purposes on almost any topic in the United States and in many other countries. Academic, governmental, and commercial organizations implement telephone surveys routinely for marketing, policy, media, public opinion, value assessment, demographic background, impact assessment, and many other purposes. The telephone survey has replaced the face-to-face household survey as the most often implemented survey technique because of improved probability sampling techniques; enhanced coverage, with virtually every household having a telephone; tried and tested techniques of question wording and questionnaire construction; improved quality control over the data-gathering process; and advances in telephone technology. These improvements, coupled with the demand of the consumers of SURVEY research for almost immediate retrieval of results, the lower costs of telephone research, and declining response rates to face-to-face surveys, have led to the prominence of surveys by telephone.

ADVANTAGES

Several factors contribute to the popularity of surveys by telephone. First, the use of the telephone

is reinforced by social norms that “demand” a person answer a ringing telephone within 3 to 4 rings, and that the person calling (i.e., the interviewer) be the person to terminate a conversation. Second, telephone surveys are cheaper than those conducted face-to-face and cost about the same as mail surveys because of lower field and personnel costs. Third, the time it takes to complete a telephone survey and get the results to the researcher or sponsor is considerably less than for any other type of survey. In fact, “instant polls,” often used by political figures seeking a public’s response to a policy proposal, can produce results within 24 hours of the first call. Fourth, CONFIDENTIALITY of respondent identity and response is enhanced because the telephone interviewers know fewer respondent identifiers, usually only a telephone number, than do interviewers in other surveys. Fifth, the effects of interviewer characteristics on response are less than those for face-to-face interviews because there is only voice contact, and the race or age of either interviewer or respondent is not generally identifiable. Therefore, telephone responses are less susceptible to socially desirable response patterns because of these physical features. Sixth, telephone surveys, particularly with the establishment of telephone centers that operate computer-assisted telephone interviewing (CATI), provide the opportunity for quality control over the entire data-gathering process because supervisors can provide immediate feedback to interviewers. In addition, these facilities almost always have the capacity for monitoring the interview via a satellite computer terminal. Seventh, coverage is no longer a problem because virtually 100% of households have telephone access, and RANDOM-DIGIT DIALING (RDD) procedures make it possible to access a telephone number without depending on a SAMPLING FRAME that lists addresses or names. Established techniques for sampling within households, such as the last/first birthday or household enumeration procedures, make it possible to continue the probability selection of a respondent. Finally, it is possible to ask questions on almost any given topic, sensitive or not. Because the telephone is such a standard means of communication, and because most countries can be characterized as “interview societies” where the interview, often by telephone, is the universal mechanism of inquiry, respondents are used to providing information to others, even to those they do not know.

DISADVANTAGES

Although many factors promote the use of telephone surveys, there are also detracting elements. First, it is easier to hang up on a telephone conversation than it is to shut the door in someone’s face. Thus, refusal is easier, making response rates to telephone surveys slightly lower than those obtained in the face-to-face survey. Refusal conversion is possible, but this is time-consuming and expensive and may produce only a 15% to 20% success rate. Second, the length of a telephone interview is limited to about 20 to 25 minutes, whereas face-to-face interviews can often go much longer and thus enable the inquiry to go into greater depth and detail. Although it is possible to ask OPEN-ENDED QUESTIONS in the telephone survey, recording the responses and probing for more depth are more difficult. Thus, telephone survey questionnaires tend to rely on structured, CLOSED-ENDED QUESTIONS because they seem easier to answer via telephone than are open-end, unstructured items. Third, because the interviewer and respondent are not in physical presence, the telephone interviewer cannot observe physical surroundings or obtain nonverbal cues, but can depend only on voice quality. Nor can the telephone interviewer use visual aids, as is possible with the face-to-face interview. Fourth, developments in telephone technology have certainly made it easier to conduct surveys, but recent innovations have contributed to making this task more difficult. Almost all households have either an answering machine, call waiting, call identifier, or modem access to the household line, each of which presents an obstacle to reaching a respondent. Prospective respondents almost never return a call to interviewers in response to a message left on an answering machine or interrupt a telephone conversation to take a call from a survey organization. Callbacks in these circumstances add to total survey cost. Fifth, legitimate surveys by telephone are compromised by the public’s generally negative view of surveys or inquiries by telephone as the result of annoying telemarketing or sales pitches that come after the caller has used the idea of a survey to get a “foot in the door” to a household in order to sell a product. This is also called “sugging.” Push polls or political propagandizing disguised as legitimate polling, and “frugging,” or fundraising under the guise of surveys, also contribute to this climate, promoting routine refusal of any inquiry, including the most legitimate surveys.

SAMPLING

The development of the RDD sampling technique made it possible to access almost every household in a country or community that had a telephone, whether the number was listed or not in the local telephone directory. Because, in some communities, up to 40% of all household numbers were unlisted by choice or because the number had not yet been entered in the printed directory, it was necessary to find a procedure that overcame the unlisted problem. RDD depends on a listing of numbers in use by exchange. Thus, it is possible to know which exchanges (e.g., 735, 542) are in use in an area, as well as the range of four-digit numbers in use in each exchange. Only 10,000 numbers are possible in each exchange. Using a random-number procedure, the appropriate set of numbers can be generated and called. One problem with this technique is the inability to screen out commercial numbers or cell phone numbers, thus making it more difficult to reach an eligible household. Also, the fact that many households now have more than one phone makes the probability of the selection of these respondents greater than those with one phone. Finally, even though exchanges have a geographical distribution over a city or country, they cannot be matched to smaller housing developments or neighborhoods. Locating eligible households in smaller geographic areas makes sampling more difficult and expensive, even with the ability to purchase custom-designed samples from commercial sampling vendors.

CONCLUSION

Issues of question order, question wording, response patterns, establishment of rapport, question complexity, and coding are similar regardless of type of survey. There is considerable research on these and other issues to the point that tested techniques of questionnaire construction and question writing are in place for telephone surveys and serve to reduce, at least, variations in response due to these factors. Data quality for telephone surveys is on par with that of other types and even improved as the result of lessened "demand" factors attributable to the physical presence of the interviewer. As indicated above, quality control of the interview process, and ultimately of data quality, is improved with the use of CATI systems, where questionnaires are presented to interviewers in an automated and controlled sequence of questions, reducing the potential of interviewer error.

Pencil-and-paper versions of telephone interviewing have all but disappeared in favor of the use of CATI systems. Although the latter are expensive, they provide almost instant results, the enhanced ability to schedule callbacks on a random schedule, and the ability to control the distribution of the sampling pool. Telephone surveys are now being used as part of "mixed mode" research, where several data-gathering techniques are used in the same research project. This makes it possible to increase response rates and overall sample coverage. Conducting surveys is now big business in almost every country of the world, with surveys by telephone a prominent component of that enterprise.

—James H. Frey

See also COMPUTER-ASSISTED PERSONAL INTERVIEWING

REFERENCES

- Bickman, L., & Rog, D. J. (Eds.). (1998). *Handbook of applied social research methods*. Thousand Oaks, CA: Sage.
- Frey, J. H. (1989). *Survey research by telephone* (2nd ed.). Newbury Park, CA: Sage.
- Frey, J. H., & Oishi, S. M. (1995). *How to conduct interviews by telephone and in person*. Thousand Oaks, CA: Sage.
- Lavrakas, P. J. (1993). *Telephone survey methods* (2nd ed.). Newbury Park, CA: Sage.
- Lavrakas, P. J. (1998). Methods for sampling and interviewing in telephone surveys. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 429–472). Thousand Oaks, CA: Sage.

TESTIMONIO

Testimonio is generally defined as a first-person narration of socially significant experiences in which the narrative voice is that of a typical or extraordinary witness or protagonist who metonymically represents others who have lived through similar situations and who have rarely given written expression to them. Testimonio is then the "literature of the nonliterary" involving both electronic reproduction, usually with the help of an interviewer/editor, and the creative reordering of historical events in a way that impresses as representative and "true" and that often projects toward social transformation. In the 1970s and 1980s, many testimonios appeared, especially throughout Latin America, including narratives of indigenous representatives and labor and guerrilla leaders. In

the 1990s, testimonios shifted toward representatives of nongovernmental organizations and broader social movements that favored human, women's, and gay rights, as well as ecological and antiglobal movements. *I, Rigoberta Menchú* became the most famous of many testimonios, and the recent attacks against this text have problematized testimonio as a viable expressive form.

Testimonio suggests testifying or bearing religious or legal witness, and the overall narrative unit is usually a life or a significant life experience. Many critics have seen the form as an expression of those marginalized by capitalist modernization. Testimonio was also seen as involving the intervention or mediation of a lettered intelligentsia. Elzbieta Sklodowska (1992) gave the most nuanced treatment of testimonio as a literary form, but John Beverley (1989; Beverley & Zimmerman, 1990) presented the most influential interpretation of testimonio's social dimensions. For Beverley and Zimmerman (1990), "Testimonio [was] usually a novella-length first person narrative recounted by the protagonist or witness to the events recounted" (p. 173). Furthermore, testimonio signaled an overall transformation in social and literary formations and even in transnational modes of production (p. 178).

Noting that testimonio-like texts had long been at the margin of literature, Beverley (1989) argued that testimonio was marked by its conflictive relation with established literary-aesthetic norms and with the institution of literature itself. Testimonio could be defined as "a nonfictional, popular-democratic form of epic narrative, since the narrative has a metonymic function which implies that any life so narrated can have a kind of representativity" evoking an implicit "polyphony of other possible voices, lives and experiences" (Beverley & Zimmerman, 1990, p. 175).

The presence of the "real," popular voice in testimonio is, in varying degrees, an illusory or indeterminate "reality effect" that can be readily challenged (cf. David Stoll's, 1999, effort to discredit Rigoberta Menchú and her book), but the effect's persistence helps to explain testimony's special force. A testimonial text simultaneously represents certain perspectives and questions supposed verities, but it is never "the truth."

Testimonio has declined in recent years, along with the social upheavals that produced the key examples. However, as an expression of gender and other identity-shaping differentials, testimonio on its own, and in its use in novels, poems, and plays, will long

remain part of Latin American and global literature as a discursive mode available to individuals related to subaltern sectors and social movements.

—Marc Zimmerman

REFERENCES

- Beverley, J. (1989). The margin at the center: On testimonio. *Modern Fiction Studies*, 35(1), 11–28.
- Beverley, J., & Zimmerman, M. (1990). *Literature and politics in the Central American revolutions*. Austin: University of Texas Press.
- Menchú, R. (1984). *I, Rigoberta Menchú: An Indian woman in Guatemala*. Ed. and intro. Elisabeth Burgos-Debray. Trans. Ann Wright. London: Verso.
- Skłodowska, E. (1992). *Testimonio hispanoamericano: Historia, teoría, poética*. New York: Peter Lang.
- Stoll, D. (1999). *Rigoberta Menchú and the story of all poor Guatemalans*. Boulder, CO: Westview.
- Zimmerman, M. (2001–2002). Rigoberta Menchú, David Stoll, subaltern narrative and testimonial truth: A personal testimony. *Antípodas: Journal of Hispanic and Galician Studies*, 13/14, 103–124.

TEST-RETEST RELIABILITY

Individual responses or observed scores on a measure tend to fluctuate because of temporary changes in respondents' states. Test-retest RELIABILITY provides an estimate of the stability (or instability) of the responses on the same measure over time. In other words, it assesses the amount of measurement ERRORS resulting from factors (e.g., maturation, memory, stressful events) that occur during a period of elapsed time. The procedure to ESTIMATE test-retest reliability is as follows. The same measure is administered to one group of participants at two different times with a certain amount of elapsed time between administrations. The two sets of scores on the measure are then correlated using the following formula:

$$r_{XX'} = \frac{\sum (X_i - \bar{X})(X'_i - \bar{X}')}{n s_X s_{X'}}$$

where X and X' represent the two administrations of the same measure, s_X and $s_{X'}$ are STANDARD DEVIATIONS of X and X' , and n is the number of respondents in the sample. This CORRELATION coefficient represents the estimate of test-retest reliability and is also referred to as the coefficient of stability.

Theoretically, a test-retest reliability estimate can range from 0 to 1, although it is not unlikely to observe negative values for measures prone to fluctuating responses. In general, there is no established minimum value that indicates an acceptable level of test-retest reliability. Additionally, the expected size of the test-retest estimate may depend upon the nature of the construct being assessed. For instance, it is unrealistic to expect high test-retest reliability for a measure that assesses transitory properties (e.g., mood). Therefore, test-retest reliability estimates tend to be higher for aptitude tests or personality measures, considered to be assessing more stable constructs, when compared to attitudinal measures or polls.

The test-retest approach is an appropriate reliability estimation technique when (a) little or no change in the characteristic being assessed is assumed to occur between the two administrations, and (b) the test material will not affect responses to subsequent presentations of the stimuli. Regarding the first circumstance, the amount of elapsed time between test administrations is an important but often neglected issue. The amount of time should be long enough to avoid any effects from memory or practice but should not be so long that the participant's level of the characteristic of interest may actually change. As a heuristic, the determination of time lapse between the two administrations should be guided by the construct's theory. According to the development of intelligence, adults' intelligence scores are not expected to change drastically between the ages of 20 and 25. High test-retest reliability (e.g., 0.80) with a 5-year time lapse provides stronger evidence of VALIDITY than the same test-retest reliability with a 2-week time lapse. On the other hand, a presumably reliable intelligence measure for infants may have very small test-retest reliability if the period of time lapse is longer than a year.

Regarding the second circumstance, research has indicated that a test-retest reliability estimate may be affected by psychological factors such as practice effects for cognitive tests or carry-over effects for attitudinal measures. For instance, people taking the same cognitive ability test may understand the problems better because of the second reading, or people may respond consistently to survey questions simply because they remember how they responded to the items previously. These BIASES could inflate the reliability estimate, as in the case where participants respond in an artificially consistent manner, but they could also result in a lower test-retest reliability

estimate, such as in the case where participants score better on a cognitive ability test the second time, causing less consistency between test scores.

—Autumn D. Krauss and Peter Y. Chen

REFERENCES

- Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. New York: Harcourt Brace Jovanovich.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

TETRACHORIC CORRELATION. See SYMMETRIC MEASURES

TEXT

The term *text* is a beguilingly short word that has become extremely elastic in meaning. Historically, its initial associations were with written forms of communication, especially those from biblical and legal sources, carrying with them a sense of authority and authoritative meaning. This sense was carried over into the idea of an author considered as the sole basis of meaning in a text. Text-based criticism thus found its rationale in locating and commenting on the author's intended meaning. Such a conception of a text has been radically overhauled in several major ways. What is now questioned is not only whether an author's intended meaning is knowable, but also whether we can speak of a text's single, final, or "true" meaning. Readers are held to construe their own meanings from a text, and these are always context-dependent. At one extreme, this leads to the reductions of historicism and cultural RELATIVISM, but rejection of the narrow focus of authorship studies, and the notion of a text's univocal meaning, is well-founded. Texts are plural and variable in their meanings, both within and across time. Just as a text is constructed in terms of particular codes and conventions, so also is it interpreted and understood; hence, its inherent semantic instability. A text cannot mean just what we take it to mean, though, for we must always take our interpretative cue

from what is set forth on the page. Yet it is not just written or word-processed communication that is now in the frame, for the term *text* has been extended to such diverse cultural forms as television, film and photography, popular song and music, sport and fashion, advertisements, household appliances, and architecture. Exclusive attention to the textuality of such forms can lead to formalist and antirealist positions, as in certain branches of POSTSTRUCTURALISM, which hold that texts have no reference but to themselves and other texts, in an endless play of signification and intertextuality. It is more generally accepted that the relations between texts and their social, cultural, and historical contexts are mediated by the narratives, genres, and discourses in which they operate and in terms of which they derive their wider meanings. While all the major disciplines in the human and social sciences have now had to negotiate the many implications of this, the adequacy of interpretation and understanding based only—or primarily—on textual analysis has just as generally been questioned. Analyzing texts is indispensable in developing a broad understanding of what is given to us as we attend to different forms of communication—whether these reach us through page, screen, or speaker—but it cannot tell us everything that is involved in their production and consumption. Broad understanding requires other methods of study alongside textual analysis. The danger of a self-confined textual analysis lies in assuming that it can embrace all that contributes to a text's construction and interpretation, as if all of that can, indeed, be "read off" from the text alone. This raises a further danger. It is one thing to say that the way a text, in its extended sense, is structured and assembled mediates what it says of the world, but quite another to assert that the world is, always and everywhere, only constituted in texts and text-analogues. The danger is then that of textualizing the world. This means seeing the world exclusively in terms of the text, outside of which nothing exists. To the contrary—we make the world meaningful only within cultural texts, but their significance lies in the references they establish to the world that exists beyond them.

—Michael J. Pickering

REFERENCES

- Hoey, M. (2001). *Textual interaction*. London: Routledge.
- Lehtonen, M. (2000). *The cultural analysis of texts*. London: Sage.
- Titscher, S., Meyer, M., Wodak, R., & Vetter, R. (2000). *Methods of text and discourse analysis*. London: Sage.

THEORETICAL SAMPLING

The relation between data and theory rests at the heart of empirical social science, and SAMPLING helps to define that relationship. RANDOM SAMPLING are used to generate DATA for purposes of verifying hypotheses derived from existing THEORY. In contrast, theoretical sampling, which is contained within the GROUNDED THEORY approach (Glaser & Strauss, 1967), is an analytical process of deciding what data to collect next and where those data should be found. It is a version of ANALYTIC INDUCTION in that an emphasis is placed on the analyst jointly collecting, CODING, and analyzing data for purposes of developing theory as it systematically emerges from data.

The central issue in this procedure centers on the theoretical purposes for collecting new data as well as deciding what kinds of data are needed for those purposes. This process goes hand in hand with the CONSTANT COMPARISON method, in which categories, situations, and analytical dimensions are continuously searched for during the research process. By combining theoretical sampling and constant comparison, theory that is tightly integrated and tied to its supporting data can be developed more efficiently.

Integration of a developing theory can be provided through several means, but emphasis is placed on two primary mechanisms. The first is selecting new data sites that specifically address provisional hypotheses grounded in data. Maines (1992) illustrates this process in his study of New York subway riders. Attempting to explain rider distributions, he first collected observational data from riders in subway cars, which led to a hypothesis suggesting that observational data should be collected from riders' waiting positions on subway platforms, which then led to a new hypothesis suggesting that observational data be collected on riders' routes of movement through the cars, which finally led to a hypothesis requiring interview data comparing riders who remained in the car they entered with those who walked to another car before taking a seat. At each stage of the research project, provisional theories were constructed that explained existing data, and qualities of those provisional theories suggested the kinds and sources of new data required to improve the explanatory capacity of the developing theory.

The second mechanism is CODING. Strauss (1987, Chap. 3) identifies three types of coding procedures that are linked to theory development. The first is *open coding*, in which data are explored for purposes of discovering categories and central themes. This process typically occurs early in the research process. The second, *axial coding*, focuses on the identification of core categories and their properties, and theorizing the conditions under which these properties (or dimensions) would occur as well as their relationships to other categories and their properties. The third procedure is *selective coding*, which entails the more theoretically purposeful specification of relationships among categories and subcategories.

Soulliere, Britt, and Maines (2001) demonstrate how combining theoretical sampling and coding for core categories is compatible with conceptual modeling. Focusing on the category of "Professionalism," they show how a concept-indicator model can be respecified on the basis of new data and presented as a concept-concept model. This modeling process is an explicit application of the grounded theory tenet of moving theory from concrete to more abstract and generic formulations.

—David R. Maines

REFERENCES

- Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Chicago: Aldine.
- Maines, D. R. (1992). Theorizing movement in an urban transportation system by use of the constant comparative method in field research. *Social Science Journal*, 29, 283–292.
- Soulliere, D., Britt, D. W., & Maines, D. R. (2001). Conceptual modeling as a toolbox for grounded theorists. *Sociological Quarterly*, 42, 233–251.
- Strauss, A. (1987). *Qualitative analysis for social scientists*. New York: Cambridge University Press.

THEORETICAL SATURATION

Theoretical saturation is the phase of qualitative data analysis in which the researcher has continued sampling and analyzing data until no new data appear and all concepts in the theory are well-developed. Concepts and linkages between the concepts that form the theory have been verified, and no additional data are needed. No aspects of the theory remain hypothetical. All of the conceptual boundaries are marked, and

allied concepts have been identified and delineated. NEGATIVE CASES must have been identified, verified, saturated, and incorporated into the theoretical scheme.

Underlying theoretical saturation is the notion of *theoretical sensitivity*. The main assumption of theoretical sensitivity is that data analysis is *data driven*. That is, categories cannot emerge until they "earn their way" into the theoretical scheme (Glaser, 2002). This means that some categories that are almost always included in a study, such as gender, cannot assume a place in the study until data analysis reveals that the constructs demand to be included. Once these constructs have appeared, sampling must continue until they are saturated.

The term *theoretical saturation* is used primarily in connection with GROUNDED THEORY, but may be applied to any qualitative research that has the end goal of developing a qualitatively derived theory. Qualitatively derived theory differs from theory used in QUALITATIVE RESEARCH—whereas theory in QUANTITATIVE RESEARCH is based on and derived from previous inquiry (and the proposed project is an extension of that work), theory in qualitative inquiry is developed systematically from data. Although the results remain a *theory* (in that it is a conceptual abstraction of those data), qualitatively derived theory differs from theory in quantitative inquiry in that it more closely resembles reality. Qualitatively derived theory is the product or outcome of inquiry, not a conceptual framework to be tested, as in quantitative inquiry. In qualitatively derived theory, concepts must be well developed and their linkages to other concepts clearly described (Morse, 1997). Criteria for evaluating the emerging theory are that it must be comprehensive, parsimonious, and complete; it must have grab; and it must be useful (Glaser, 1978).

An example of a theory that is developed using theoretical saturation is the classic monograph by Anselm Strauss and his colleagues (1984). In this grounded theory, they describe many aspects of individuals' experiences with chronic illness, including the trajectory of illness and hospitalization, the impact on the family, and the use of community services such as self-help groups. They write of the basic strategy, for instance, not as symptom control or staying alive, but as *normalizing*. In this research, the process of making their lives (and their families' lives) as normal as possible depends not only on the arrangements, but also on "how intrusive are the symptoms, the regimens, the knowledge that others have the disease and of its

fatal potential” (Strauss et al., 1984, p. 79). If these illness factors become intrusive, then people have to “work very hard at creating some semblance of normal life for themselves” (p. 79). The description of what Strauss et al. term “abnormalization and normalization of life” is a excellent illustration of the richness and the complexity of qualitatively derived theory that may be attained when the investigators have used the principles of theoretical sensitivity during data collection and analysis.

—Janice M. Morse

REFERENCES

- Glaser, B. G. (1978). *Theoretical sensitivity*. Mill Valley, CA: Sociology Press.
- Glaser, B. G. (2002). Grounded theory and gender relevance. *Health Care for Women International*, 23, 786–793.
- Morse, J. M. (1997). Considering theory derived from qualitative research. In J. Morse (Ed.), *Completing a qualitative project: Details and dialogue* (pp. 2163–2188). Thousand Oaks, CA: Sage.
- Strauss, A. L., Corbin, J., Fagerhaugh, S., Glaser, B., Maines, D., Suczek, B., & Weiner, C. (1984). *Chronic illness and the quality of life*. St. Louis, MO: Mosby.

THEORY

There can be few terms in the social sciences whose usage is more diverse, and often vague, in meaning than “theory.” There are at least three different contrasts involved: with practice, with fact, and with evidence. In opposition to “practice,” “theory” often refers to stated principles rather than to what is actually done. There is generally an evaluative implication here, and it can go in either direction. Theory may be held up as an ideal that shames practitioners and/or as something that, if followed, can transform practice. Alternatively, theory may be dismissed as unrealistic or even as mere rationalization. More balanced interpretations are also possible: Here, theory consists of normative principles that provide a general guide for practice, without supplying detailed instructions.

The second contrast, with fact, can also have an evaluative slant in different directions. Theory may be dismissed as speculative and as something to be avoided. This is an attitude found among some EMPIRICISTS. Alternatively, “fact” may be taken to refer to historically constituted appearances that present themselves

as universal and that are therefore systematically misleading. Given this, the role of theorizing is to penetrate factual appearances in order to identify underlying, generative structures, and thereby to show that fundamental social change is possible. This meaning of theory is largely derived from Marxism and CRITICAL THEORY.

The third meaning of theory involves a contrast with evidence or DATA. Here, theories are factual, in the sense of being neither valuational nor speculative. Nevertheless, there is considerable variation. At one end of the spectrum are theoretical frameworks or approaches; examples include BEHAVIORISM, functionalism, rational choice theory, and SYMBOLIC INTERACTIONISM. At the other extreme are specific sets of explanatory principles relating to particular types of phenomena and the conditions under which they occur. An example is differentiation-polarization theory in the sociology of education. This claims that the differentiation of students—for example, through streaming, tracking, banding, or setting—generates a polarization in their attitudes toward school, with those ranked highly becoming more positive and those ranked at the bottom more negative. Of course, other theories lie somewhere between these two extremes. One example would be labeling theory in the sociology of deviance. In specific terms, this claims that, contrary to common belief, labeling and punishing offenders can increase, rather than decrease, the chances of their offending again in the future. However, labeling theory also serves as a theoretical framework, one that emphasizes the importance of studying the labelers as well as the labeled, and requires a suspension of conventional views about what is right and wrong.

Qualifiers are sometimes attached to theory. These may label a particular theoretical approach, such as “critical” or “rational choice” theory. Others mark out a contrast in the level of abstraction involved, as in the case of “grand” and “middle range” theory (Merton, 1968, pt. 1). Other qualifiers are methodological, referring to how theory is produced (e.g., GROUNDED THEORY).

—Martyn Hammersley

REFERENCES

- Hammersley, M. (1995). Theory and evidence in qualitative research. *Quality and Quantity*, 29, 55–66.
- Merton, R. K. (1968). *Social theory and social structure* (3rd ed.). New York: Free Press.

THICK DESCRIPTION

The anthropologist Clifford Geertz popularized the term *thick description* among social scientists, having borrowed it from the philosopher Gilbert Ryle (Geertz, 1973; Ryle, 1971). Both writers reject “thin” descriptions, or those that seek to reduce human actions to observable behavior. The need for thick description stems from the fact that actions are defined by cultural meanings, not by their physical characteristics. For instance, a nervous tick can be contrasted with a wink: The overt behavior may be the same, but the second is an action whereas the first is not. And a wink is different from a parody of a wink, and from somebody practicing winking. As these examples make clear, actions are meaningful, and meaning depends on context—on who is involved in doing what, when, how, why, in response to what other actions or events, with what consequences, and so on. Therefore, Geertz argues, sound description must include the web of meanings that is implicated in the actions being described.

He contrasts thick descriptions not just with BEHAVIORISTIC records of human actions but also with STRUCTURALIST accounts that abstract meanings from their local contexts. He writes that in anthropological work, “theoretical formulations hover so low over the interpretations they govern that they don’t make much sense or hold much interest apart from them” (Geertz, 1973, p. 25). The task in thick description is both to capture the complexity of particular events and to indicate their more general cultural significance. Probably the best known example of a thick description is Geertz’s article “Deep Play: Notes on a Balinese Cockfight” (in Geertz, 1973), in which he uses detailed analysis of a single witnessed event to illuminate distinctive features of Balinese culture.

Geertz proposes that producing thick descriptions should be the primary aim of anthropological research. He insists that such interpretive anthropology is no less a science than physics or chemistry, but that it is a different kind of science. At the same time, he compares the task of the anthropologist with that of the literary critic. In these terms, he suggests, ETHNOGRAPHY is analogous to trying to construct a reading of a manuscript, one that is “foreign, faded, and full of ellipses, incoherencies, suspicious emendations and tendentious commentaries” (Geertz, 1973, p. 10).

The term *thick description* has been adopted by many QUALITATIVE RESEARCHERS as a way of signaling

the characteristic orientation of their work. And the concept has also been used to address the question of how the findings of qualitative studies can be generalized. Thus, Guba and Lincoln argue that thick description is required for readers of qualitative research to be able to assess the transferability of findings to other contexts (Guba & Lincoln, 1985; see also Gomm, Hammersley, & Foster, 2000).

—Martyn Hammersley

REFERENCES

- Geertz, C. (1973). *The interpretation of cultures*. New York: Basic Books.
- Gomm, R., Hammersley, M., & Foster, P. (2000). Case study and generalisation. In R. Gomm, M. Hammersley, & P. Foster (Eds.), *Case study method: Key issues, key texts*. London: Sage.
- Guba, E. G., & Lincoln, Y. S. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.
- Ryle, G. (1971). Thinking and reflecting. In G. Ryle, *Collected papers, Volume 2*. London: Hutchinson.

THIRD-ORDER. See ORDER

THRESHOLD EFFECT

The threshold effect takes place in instances where the relationship between two variables appears to seriously change on arriving at a certain value of one or both variables. MODELING the threshold effect is of paramount importance for the analyst in order to describe correctly the relationship between the two variables.

An illustration of the threshold effect is the relationship between the economic system of a country and its level of democratic performance. Assume that one can MEASURE the extent to which a country is capitalist. The measure can be at the ordinal level, as in the “economic freedom” indexes that are put forth by organizations such as the Fraser Institute. It can also be an INTERVAL measure, such as government spending as a percentage of gross national product. With capitalism as the INDEPENDENT VARIABLE, further assume that democracy as the DEPENDENT VARIABLE can be measured on a 0–100 scale, with low scores indicating

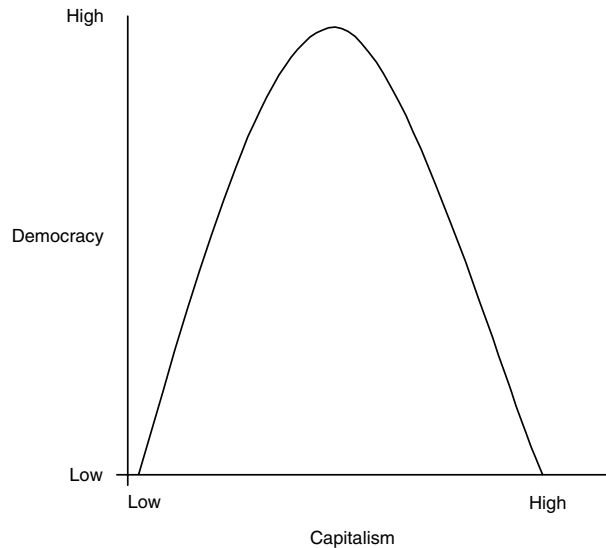


Figure 1 Parabolic Relationship of Capitalism to Democracy

the country is less democratic (i.e., less respecting of civil rights and liberties), and high scores revealing a more democratic country (i.e., more likely to hold free and fair elections). Those who believe that there is a threshold effect to the relationship between capitalism and democracy believe, in theory, that capitalism is useful only to a point, whereupon the more capitalist the country, the less democratic it will be. The theory here is that highly capitalist countries, with lower levels of government intervention in the economy, will not be able to provide the social services, the social safety net of welfare spending, or law enforcement that allows for a full expression of democracy in the country.

Empirical studies using OLS regression modeling techniques reveal that, indeed, there is a STATISTICALLY SIGNIFICANT threshold effect of capitalism on democracy, beyond which capitalism has a negative impact on democratic performance in the country. Graphically, the threshold effect is one of an upside down U-shaped curve, with the economic freedom that comes from increasing levels of capitalism in the lower range also increasing democratic performance, up to a point. Beyond the point, the relationship reverses, from positive on the graph to negative on the graph, with higher levels of capitalism actually decreasing democratic performance and forming the right-hand side of the upside down U-shaped curve.

The correct way to model this threshold effect is to take the square of the independent variable—in this case, capitalism—and enter it as a second independent

variable in the model. If there is a threshold effect, then both the linear term and the squared term will achieve conventional levels of statistical significance ($p < .05$, two-tailed test). The mathematical way to calculate the independent variable value at which the threshold effect takes place, with the SLOPE changing from positive to negative on the graph (called the “tipping point”) is to use the formula $-b_1/2b_2$, where b_1 is the OLS REGRESSION COEFFICIENT for the linear term of the independent variable, and b_2 is the OLS regression coefficient for the squared term of the independent variable.

Another illustration of a threshold effect is described by Breen (1996, p. 50). Currency exchange rates are directly observed only when they fall within an upper and lower range of values; in other words, an upper and lower threshold. If the analyst seeks to explain currency exchange rates, then this example is also one of a censored regression model. The analyst can estimate censored regression models using a Tobit estimator.

No matter the example, the analyst will go astray in interpreting the results of parameter estimation without the proper modeling of threshold effects.

—Ross E. Burkhart

REFERENCES

- Breen, R. (1996). *Regression models: Censored, sample selected, or truncated data* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–111). Thousand Oaks, CA: Sage.
- Brunk, G. G., Caldeira, G. A., & Lewis-Beck, M. S. (1987). Capitalism, socialism, and democracy: An empirical inquiry. *European Journal of Political Research*, 15, 459–470.
- Burkhart, R. E. (2000). Economic freedom and democracy: Post-cold war tests. *European Journal of Political Research*, 37, 237–253.

THURSTONE SCALING

Thurstone scaling is a covering term representing a series of three techniques developed by psychometrician Louis L. Thurstone during the 1920s. These scaling methods are commonly known as

1. The Method of Equal Appearing Intervals
2. The Method of Successive Intervals
3. The Method of Paired Comparisons

Beneath all of these scaling techniques was a common perspective on the measurement of attitudes—the Law of Comparative Judgment (Thurstone, 1927).

Imagine that you wish to compare the weights of multiple objects and have no modern scale or set of standard weights. Rather than despair about an inability to judge weights precisely, one can estimate the weights of objects relatively—ordering them from lightest to heaviest. This could be done by comparing each object against the others one at a time, judging them in pairwise comparison. Place one object on either side of a balanced beam and identify the heavier object. A full set of comparisons should yield a set of objects ordered by weight. Thurstone was the first to realize and implement a system to do the same thing with psychological rather than physical objects.

A scaling methodology designed to assign numbers to the value or intensity of psychology objects must reflect the discriminations made by individuals judging multiple stimuli. The bottom line underlying the scaling of comparative judgments is that objects that are easily discriminated should be placed further apart than objects that are difficult to distinguish.

The Method of Paired Comparisons involves explicit judgment by all subjects of each combination of objects. This task becomes formidable as the number of objects increases, because the number of paired comparisons rises geometrically. That is, assuming a pairing of A and B is the same as a comparison of B and A, the number of comparisons required equals $N(N - 1)/2$. Four objects will be paired in 6 ways, 5 in 10 ways, 10 in 45 ways. If order matters, the number increases even more rapidly: $N(N - 1)$. Four objects will generate 12 comparisons, 5 call for 20, and 10 require 90. The ordering of a large number of objects becomes unwieldy.

The Method of Equal Appearing Intervals is a technique developed to capture individual assessments of attitude objects without the necessity of capturing all paired comparisons. Judges are employed to calibrate a set of items, which is then administered to the study population.

Begin with a large set of statements reflecting diverse opinions toward an object. Present these statements to a panel of judges, each of whom determines the degree of support for (or opposition to) the object under study by placing the statement in one of 11 ordered piles ranging from “most favorable” to “neutral” to “least favorable.” These initial judgments

provide the basis for the selection of items for the final scale.

Items are ranked in terms of their average support score (MEDIAN) and the variation in support scores (INTERQUARTILE RANGE). Those items that are inconsistently rated by the judges are dropped, and a final scale of 5 to 20 items is created by selecting items covering the range of support and approximately evenly spaced (as presented by the median assessment of the judges).

Equal-appearing intervals scales tend to truncate variation at their extreme points. The Method of Successive Intervals is developed as a check on the assumption made in the procedure for equal interval scaling that all intervals are, in fact, equal. Successive Interval methods allow for unequal intervals to capture extreme positions.

Thurstone scaling methods are limited in several respects:

- *Assumption of unidimensionality.* These scales do not account for the multidimensional nature of attitudes, instead constraining responses to conform to unidimensional structure.
- *Labor intensive.* In exploring attitudes about a stimulus object, the researcher must first create an extensive bank of items representing a range of opinions about the object. The creation of these scales involves considerable work for judges in shifting through items to build final scales.
- *Assumed similarity between judges and subjects.* Judges drawn from populations distinct from those of the focus of study may produce scales that do not apply to the study population.

As a consequence of the weaknesses underlying Thurstonian methods, these scaling techniques are of historical interest more than practical use today.

—John P. McIver

REFERENCES

- Edwards, A. L. (1957). *Techniques of attitude construction*. New York: Appleton-Century-Crofts.
- McIver, J. P., & Carmines, E. G. (1980). *Unidimensional scaling*. Beverly Hills, CA: Sage.
- Scott, W. A. (1968). Attitude measurement. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology*. Reading, MA: Addison-Wesley.

- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433–451.
- Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology*, 17, 446–457.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273–286.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 23, 529–554.
- Thurstone, L. L. (1929). Fechner's Law and the Method of Equal-Appearing Intervals. *Journal of Experimental Psychology*, 12, 214–224.
- Thurstone, L. L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology*, 26, 249–269.
- Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitudes*. Chicago: University of Chicago Press.

TIME DIARY. See TIME MEASUREMENT

TIME MEASUREMENT

Time is a social product. All of the human life we know relates itself to rhythms derived from more-or-less regular “time givers” (which might be either natural or artificial: constellations, the sun and moon, church clocks, factory whistles, television programs, family rituals). The different time givers are embodied in different structures of authority, domination, or reciprocity, producing the diverse mixture of time structures through which each individual must navigate during the day. And each person experiences the passage of time differentially according to context and circumstances. Nevertheless, there is a consensual physicists’ time, counted as, for example, oscillations of a pendulum or of a cesium crystal. We are accustomed to measuring the durations of various activities against the passage of this clock time using DIARY methods. “Time budget surveys” involve SAMPLES of diaries allowing researchers to estimate patterns of time use over a given population and calendar period. More than 80 countries have collected nationally representative time diary studies. Szalai (1972) provides a cross-national diary study covering 12 countries; the Multinational Time Use Study now provides microlevel data for 25 countries and various recent historical periods for secondary analysis (Gershuny, 2000).

The diary method mirrors the structure of daily activities. Diary accounts are complete and continuous, so any respondent intending to suppress some activity must make a conscious effort to substitute some other activity. Diaries allow investigation of the activity sequences, or the times of the day at which activities occur. Simultaneous activities can be recorded, and researchers can use them to investigate patterns of co-presence, interdependence, and cooperation. Nevertheless, the time diary method is not unproblematical. In particular, it produces some “under-reporting of socially undesirable, embarrassing, and sensitive” behaviors, including personal hygiene, sexual acts, and bathing (Robinson & Godbey, 1997, p. 115).

There is wide variation in the design of time diary instruments (Harvey, 1999, provides general-purpose design recommendations). There are, as in all obtrusive survey methods, conflicts between the respondent burden (leading to better data but lower response rates) entailed by more extensive and detailed diaries, and the higher response rates permitted by so-called light diary instruments. (Summary results, however, seem quite stable with respect to the different methods; Harvey, 1999.) Particular problems result from daily variation in individuals’ activities. A short (e.g., single-day) diary gives a spurious impression of interpersonal variability that is, in fact, intrapersonal. A longer (e.g., 7-day) diary is more revealing but also much more demanding for respondents.

There are two alternative modes of diary administration, referred to as, respectively, “yesterday” studies, involving an interviewer-led process, and “tomorrow” studies, involving a self-completion instrument. The yesterday diary normally covers the previous day, but despite recall decay, it is possible to collect useable information after a gap of 2 or even 3 days; this may be required to overcome difficulties in interviewing over weekends or short holidays. The tomorrow self-completion diary is left behind with the respondent by the interviewer with the instruction to make records as frequently as possible through the diary period so as to maximize the recall quality. These normally cover either 1 or 2 days, but The Netherlands and the United Kingdom have reasonably successful experience with 7-day diaries. Studies sometimes use a combination of yesterday and tomorrow methods with the self-completion instrument being posted back by the respondent. In this case, the tomorrow approach shows lower and somewhat biased responses. For

tomorrow diaries, response is substantially improved if the interviewer returns to collect the diary.

A further design issue concerns the temporal framework of activity recording. Yesterday diaries often adopt an event basis of recording ("When did you wake up?" "What did you do first?" "What did you do next?"). For self-recording, it is more appropriate to use a schedule with regular time slots of 5 to 15 minutes (although slots of up to 30 minutes have been used, and the length of the slots may vary through the day and night); respondents indicate the duration of each activity on the schedule using lines, arrows, or ditto marks.

Activities may be reported according to a comprehensive "closed" list of precoded activities, or an "open" ("own words") approach. The closed approach may overconstrain respondents, and the choice of categories may, to a degree, influence the level of reporting. But the open approach produces a symmetrical problem: Respondents report at varying levels of detail that are difficult to disentangle from genuine interpersonal differences.

For much of the day, respondents are engaged in multiple simultaneous activities. The normal practice is to ask respondents to record activities in a hierarchical manner: ("What are/were you doing?" "Were you doing anything else at the same time?"). Some studies ask respondents to record all activities in a nonhierarchical manner, but respondents tend to prefer opportunity to use a narrative mode ("First I did this, and then I did that . . ."), which naturally involves the hierarchical approach. The hierarchy also simplifies the process of analysis. Although primary activities are the most frequent focus of analytical attention, it is also possible to represent combinations of activities nonhierarchically while still conserving the 1440 minutes of the day.

Other ancillary information included within the rows or columns of diary forms include co-presence ("Who else is engaged with you in this activity?") and current location (or modes of transport). Specialized diary studies also include interpretive or affective information (e.g., "Is this work?" or "How much did you enjoy this?") attached to each event or activity.

Beyond the design of the diary instrument, there are also various specialized sampling issues. Seasonal effects may be of importance; representative samples of time use require an even coverage across the year. Conventional tomorrow studies have included children down to the age of 10.

An alternative approach to the measurement of time use involves "stylized estimates" ("How much time do you spend in . . .?") as part of questionnaire inquiries. This approach is the simplest and cheapest way to collect information about how people use their time, and it is widely used (e.g., in national labor force surveys). But it has significant drawbacks. Research on the paid work time estimates deriving from this methodology suggests systematic biases, particularly overestimates of paid work time (Robinson & Godbey, 1997, pp. 58–59).

ETHNOGRAPHIC research has often involved direct observation of people and the recording of their activities by the researchers. New technology now extends the role of direct observation, including such unobtrusive devices as individual location monitoring systems and closed-circuit or Internet camera systems. Observational methods have advantages, but they are costly and labor intensive, and they may compromise personal privacy. Moreover, people can be expected to censor or distort their normal patterns when aware of continuous monitoring.

—Jonathan Gershuny

REFERENCES

- Gershuny, J. (2000). *Changing times: Work and leisure in post-industrial societies*. Oxford, UK: Oxford University Press.
- Harvey, A. S. (1999). Guidelines. In W. E. Pentland, A. S. Harvey, M. P. Lawton, & M. A. McColl (Eds.), *Time use research in the social sciences* (pp. 19–45). New York: Kluwer.
- Robinson, J. P., & Godbey, G. (1997). *Time for life*. University Park: Pennsylvania State University Press.
- Szalai, A. (Ed.). (1972). *The use of time*. The Hague: Mouton.

TIME-SERIES CROSS-SECTION (TSCS) MODELS

TSCS models aim to detect the causal processes underlying TSCS data (often called PANEL data). One can represent such data as a three-dimensional "data cube" in which cases, variables, and time points are the dimensions (see Figure 1). TSCS data sets contain more information than either (a) CROSS-SECTIONAL DATA sets comprising data on variables for cases at a single time (e.g., the front slice of the cube in Figure 1), or (b) TIME-SERIES data sets comprising data

on variables at times for a single case (e.g., the top slice of the cube in Figure 1).

Researchers sometimes compile TSCS data sets because a target population has a small number of cases. For example, any cross-sectional study of European nation states will be limited by the fact that it will have fewer than 30 cases, severely restricting one's ability to carry out multivariate statistical analyses. By collecting data for those cases at several time points rather than just one, a researcher may be able to obtain enough additional observations to facilitate better analyses.

TSCS data sets are also attractive because of their analytic potential. For example, using TSCS data, one may be able to assess whether an independent variable's effect on a dependent variable changes over time (not possible with cross-sectional data), or whether an independent variable's effect differs across cases (not possible with simple time series data). Finally, researchers frequently use TSCS data in order to control for case-stable unmeasured variables, an important type of unmeasured HETEROGENEITY, in multivariate analyses.

The analytic flexibility that TSCS data affords is not unlimited, however. To gain an advantage from pooling time series and cross-sectional data, there must be reason to assume some kind of invariance in the effects, including nonlinear and nonadditive effects, of the independent variables. These effects need to be constant across cases, across time, or both (in practice, most researchers assume both kinds of invariance). In addition, in TSCS analyses, both HETEROSKEDASTICITY, primarily seen as a potential problem for cross-sectional STATISTICAL INFERENCE, and AUTOCORRELATION, primarily seen as a potential problem for time series statistical inference, are possible threats to statistical inference.

LARGE-*N*, SMALL-*T* TSCS MODELS

Given the many analytic opportunities afforded by TSCS data, there are many possible TSCS models. Two general classes of models exist, however,

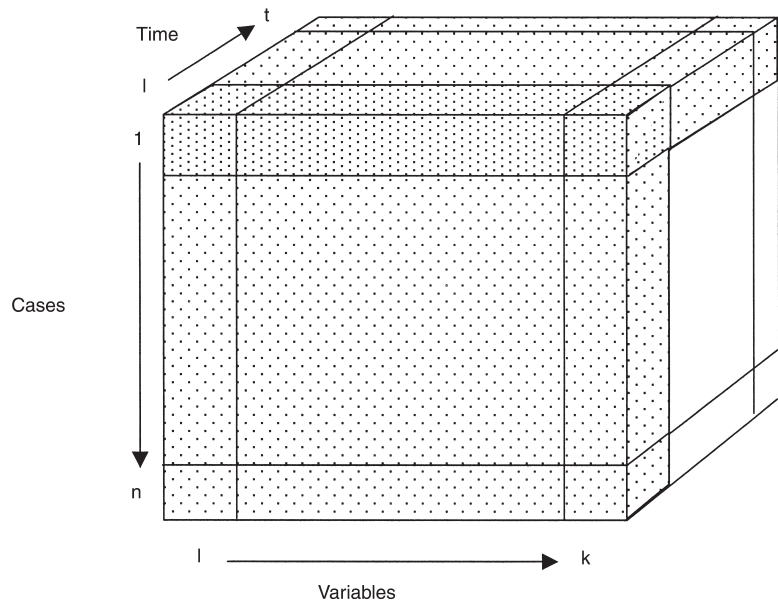


Figure 1 A Data Cube Representation of a Time Series Cross-Section Data Set

corresponding to whether the number of cases (n) substantially exceeds the number of time points (t), or vice versa. Large n panel studies, such as the National Longitudinal Study of Youth and the Panel Study of Income Dynamics in the United States, have thousands of cases and relatively few (i.e., < 15) time points. Studies using such data typically view unmeasured heterogeneity among the cases as the most important potential source of estimator bias and inconsistency, and address this problem using either fixed effects (FE) or random effects (RE) estimation procedures. Both sets of estimation procedures stem from the analytic model:

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_k X_{kit} + \alpha_i + \varepsilon_{it},$$

where i indexes cases, t indexes times, and numerical subscripts index the measured independent variables (X), with β_0 standing for the equation's intercept term. As usual, ε_{it} represents a DISTURBANCE TERM with a mean of zero, constant VARIANCE, and no ASSOCIATION with the independent variables in the equation. The α_i represent the effects of unmeasured variables that are stable across time for each case, but vary across cases. In studies where individuals are cases, for example, this term would represent stable but unmeasured variables, such as innate ability or physical attractiveness, that affect cases' values on the dependent variable.

The FE estimator for this model includes the α_i among the independent variables in the equation, either directly—as a set of $n - 1$ dummy variables representing the n cases—or indirectly—as a case-specific “time demeaning” of the measured variables in the analysis (Wooldridge, 2000, pp. 441–446). Because the α_i represent variables that are stable over time for each case, using the fixed effects estimator makes it impossible to estimate partial effects for stable, *measured* independent variables, such as gender and race/ethnicity. Most researchers consider this a small price to pay for the ability to control for the effects of myriad unmeasured stable variables, however. (The first-difference estimator—see Wooldridge, 2000, pp. 419–425—although closely related to the fixed effects estimator, is used less often, even though in some situations it may have superior statistical properties; Wooldridge, 2000, pp. 447–448.)

The RE estimator for large- n , small- t TSCS data sets assumes that α is a random variable and leaves it in the equation’s error term, thereby creating a composite error term $u_{it} = \alpha_i + \varepsilon_{it}$. This approach has three important implications. First, the RE estimator is biased and inconsistent unless the variables that α represents are unrelated to the measured independent variables in the analysis. Many researchers find it difficult to assume that α will fulfill this condition and therefore prefer FE. For example, physical attractiveness (usually unmeasured) is probably positively correlated to personal income (usually measured). Second, when case-stable variables are left in the composite error term, one must use feasible GENERALIZED LEAST SQUARES estimation techniques (FGLS) to take into account the fact that the disturbance terms for each case will be correlated across time. Fortunately, many statistical packages (e.g., SAS, Stata) include a program to carry out this FGLS procedure. Third, the inclusion of α in the composite error term makes it possible to estimate partial effects for measured independent variables that vary across cases but not across time, which is impossible when one uses FE.

If α is unrelated to the measured independent variables in one’s analysis, RE will be more efficient than FE because it estimates only one additional quantity (the variance of α) beyond the standard OLS estimates. In contrast, FE estimates an additional $n - 1$ quantities when it estimates each of the α_i . Researchers frequently use a Hausman test to determine whether α is unrelated to the measured independent variables in an analysis (Wooldridge, 2002, pp. 288–291), and

thus whether using the potentially more efficient RE estimator is justifiable.

The model above can be extended to include a term for unmeasured time effects as well as unmeasured case effects. In this case, the model becomes

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_k X_{kit} + \alpha_i + \nu_t + \varepsilon_{it},$$

where ν_t represents the effects of unmeasured variables that are constant across cases at a given time. Subject to the same advantages and disadvantages outlined above for the model containing only α_i , one can use FE, RE, or a combination of FE and RE to estimate this model.

SMALL- N , LARGE- T TSCS MODELS

Researchers analyzing TSCS data sets with small numbers of cases and many time points typically focus on heteroskedastic and autocorrelated disturbances as the chief threats to valid inference, and take advantage of the many observations per case to estimate FGLS models that adjust for these threats. At a minimum, such models incorporate casewise heteroskedasticity and a common autocorrelation pattern across cases, but many other possibilities exist. For example, one popular alternative, the Parks-Kmenta model, incorporates casewise heteroskedasticity, case-specific levels of AR1 autocorrelation, and time-specific correlations among the disturbances for different cases. Estimation of complicated disturbance structures require data on many time points, however, and Monte Carlo experiments suggest that estimation of the Parks-Kmenta model can produce seriously misleading statistical tests when the number of time points is fewer than 40 (Beck & Katz, 1995). Most econometric statistical packages include a range of alternative models for analyzing small- n , large- t TSCS data.

—Lowell L. Hargens

REFERENCES

- Beck, N., & Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, 89, 634–647.
- Kmenta, J. (1986). *Elements of econometrics* (2nd ed.). New York: Macmillan.
- Wooldridge, J. M. (2000). *Introductory econometrics: A modern approach*. Cincinnati, OH: South-Western.
- Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.

TIME-SERIES DATA (ANALYSIS/DESIGN)

Time-series data arise frequently in the study of political economy, macroeconomics, macropolitical sociology, and related areas. Time-series data are repeated observations, at regular intervals, on a single entity, such as a country. They are distinguished from CROSS-SECTIONAL DATA, which consist of observations on a sample of entities (perhaps individuals or countries) at one point in time. Although some analysts analyze a single time series (see BOX-JENKINS MODELING), most use time-series data to study the relationship between a dependent variable and one or more independent variables, as in the typical REGRESSION. For simplicity, this entry considers only a single independent variable, although everything said generalizes easily to a MULTIPLE REGRESSION. Because time-series models are a type of regression model, all of the issues related to regression models apply, including testing, SPECIFICATION, and so on. This article focuses on issues specific to the analysis of time-series data by discussing, first, some examples of time-series data; then some issues in estimation; and, finally, some issues in specification and interpretation (see also LAG STRUCTURE), with a brief discussion of further directions in the concluding section.

EXAMPLES OF TIME-SERIES DATA

Macroeconomists are the most common users of time-series data. Typically, they have monthly or quarterly observations on a variety of economic phenomena and are interested in relating, say, some policy instrument, such as the money supply, to some economic variable of interest, such as the growth of GDP. Similar analysis is undertaken by political economists and sociologists who are interested in the relationship of political variables, such as the political orientation of the coalition in power, to economic outcomes, or the relationship of economic variables, such as unemployment or inflation, to political outcomes, such as presidential popularity or the outcome of elections. Time-series data arise in many other areas, such as the study of conflict, where studies analyze the causes of defense expenditures over time. In all of these studies, analysts must be aware of the frequency of their data, be it daily, monthly, quarterly, or annually; different

models are appropriate for data of higher or lower frequency.

This entry limits itself to discussing stationary time series. Roughly speaking, a time series is stationary if its various statistical properties do not change over time; in particular, the best long-run forecast of a stationary series is its overall mean. Nonstationary series involve much more difficult mathematics; causes of nonstationarity usually involve some type of time trend. Analysts must be careful that they limit their use of stationary methods to stationary time series.

Although time-series data refer to observations on a single unit, they can be generalized to TIME-SERIES CROSS-SECTIONAL DATA, where analysts study time series obtained on a variety of units, usually having annual observations on a set of countries. Time-series cross-section data have some time-series properties, and so many issues relevant in the estimation of time-series models are also relevant. PANEL data are also related to time-series data; there, the analyst usually has only a few repeated observations on a large number of individuals. Although panel data present estimation problems related to time-series issues, the data are sufficiently different that different models and approaches must be used.

ESTIMATION

Stationary time-series data are estimated using standard regression methods. If the data meet all of the assumptions for such methods, then ORDINARY LEAST SQUARES has its usual optimal properties. But time-series data usually do not meet one critical assumption—that the observations are independent of each other (or, equivalently, that the error terms for each observation are independent). Time-series data observed at closely spaced intervals (perhaps a month) in general will be highly interdependent; data observed at intervals far apart (say, a decade) might well be independent.

Dependence has serious consequences for estimation. Fortunately, it is easy to both test for this dependence and correct for it. Failure to do so can lead to highly inefficient estimates and highly inaccurate standard errors. Although violating some assumptions may not be serious, violations of the independence of error terms assumption can easily lead to mistakes of several hundred percent. Thus, time-series analysts must be very careful to both test and correct for dependencies among observations. The typical assumption

is that observations observed one time period apart are the most related, with that relationship declining as the two observations are observed further and further apart. Letting the error term for an observation at time period t be denoted ε_t , assume a simple regression model:

$$y_t = \alpha + \beta x_t + \varepsilon_t.$$

The simplest assumption about the errors is that they follow a first-order, autoregressive (AR1) pattern, that is,

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t,$$

where the v_t are assumed to be independent of each other. ρ is then a measure of how highly correlated the errors are.

Analysts can test the hypothesis $H_0: \rho = 0$, that is, that the errors are independent, by estimating the regression model and capturing the RESIDUALS. Independence can then be assessed by the famous DURBIN-WATSON STATISTIC or, more simply, by regressing the OLS residuals on their lags, plus all other independent variables, and then assessing the significance of the coefficient in this auxiliary regression. This is the simplest example of a Lagrange multiplier test; such tests are extremely useful in the analysis of time series.

If the null hypothesis of independence is not rejected, time-series data can be estimated by OLS (assuming the other assumptions are met). But if the null hypothesis of independence is rejected, as it often will be, then OLS is seriously flawed. A good alternative is GENERALIZED LEAST SQUARES. For time-series data with AR1 errors, generalized least squares is essentially the famous Cochrane-Orcutt procedure. This proceeds in five steps:

- Use OLS on the basic regression equation.
- Use the residuals from that regression to estimate ρ (by regressing the residuals on the lagged residuals).
- Use that estimate to “pseudo-difference” the observations (i.e., take the current observation and subtract the estimated ρ times the previous observation).
- Redo OLS, estimating a new value of ρ .
- Iterate Steps 3 and 4, stopping when the estimate of ρ does not change.

This procedure provides good estimates of α and β and correct standard errors.

Often, we add last period’s observation of y , the lagged dependent variable, y_{t-1} , to the specification, giving

$$y_t = \alpha + \beta x_t + \varpi y_{t-1} + \varepsilon_t.$$

Although the Durbin-Watson test no longer works with a lagged dependent VARIABLE, the Lagrange multiplier test discussed continues to be a good test of whether the errors are independent. It is often the case that errors in the lagged dependent variable model are independent, in which case the model is well estimated by OLS. But analysts must use the Lagrange multiplier test to examine the hypothesis of independence: If the errors are dependent, then OLS is not even consistent.

If x and y are not stationary, then all standard methods fail. Roughly speaking, analysts can assess stationarity by regressing x on its first lag and then examining the estimate of the regression coefficient (and similarly for y). If the coefficient is close to one (a UNIT ROOT), then analysts must use much more complicated methods, well beyond the scope of this entry. However, one model that often works well with such data is the ERROR CORRECTION MODEL, which assumes that there is a long-run equilibrium relationship between x and y as well as a short-run relationship, so that the estimated model is

$$\Delta y_t = \alpha + \beta \Delta x_t - \phi(y_{t-1} - \gamma x_{t-1}) + \varepsilon_t.$$

The term in parentheses is the amount that the system was out of equilibrium, and ϕ is the speed of adjustment back to equilibrium. Under many conditions, this model can be estimated by OLS. However, nonstationary series are complicated, and analysts must be very careful to use appropriate methods and models in this highly technical field.

Finally, it should be noted that the above is limited to models where effects either are all felt immediately or work with a one-period lag. It is easy to generalize this to more complicated LAG STRUCTURES. Although it is often the case that one-period lags suffice for annual data, quarterly or monthly data frequently require much more complicated lag structures. Here, still, the principles of testing remain similar. There are many additional tools available, such as the correlogram, to aid the analyst in deciding how many lags to include in the model.

SPECIFICATION AND INTERPRETATION

In a simple regression, the regression slope, β , tells us that if x changes by one unit, y is expected to

change β units. This is true for simple time-series models, but such models can also have more complex specifications. Thus, for example, in the lagged dependent variable model, a one-unit change in x is associated with an initial change of β units in y , but the effect of a change in x persists over time. Suppose that x goes back to its initial value in the next period. In this period, though, the lagged y is now β units higher, so the current y is $\beta\omega$ units higher. In the next period, the lagged y is $\beta\omega$ units higher, so the current y is $\beta\omega^2$ units higher, and so forth. The graph of this function is called an impulse response function.

Alternatively, we could allow for x to change one unit and stay at that new level forever (level change analysis). How, then, does y respond? Initially, y goes up by β units. In the next period, lagged y is β units higher, but x is still one unit higher, so y is $\beta + \omega\beta$ units higher. The following period, x is still one unit higher, and y is now $\beta + \omega\beta$ units higher, so y is now $\beta + \omega(\beta + \omega\beta) = \beta + \omega\beta + \omega^2\beta$ units higher. This geometric progression continues, and so, by simple algebra, a one-unit change in x is eventually associated with a $\beta/(1 - \omega)$ -unit change in y . (This assumes that $\omega < 1$, which is a necessary condition for stationarity.) Thus, if ω is large (say, 0.9), the long-run impact of a unit change in x is 10 times the short-run impact, β . It is critical to keep this interpretation in mind. All time-series models should always be interpreted with both impulse response and level change analyses. (It is important to remember that all of the coefficients used are estimated, and so estimated effects have confidence intervals.)

Such analysis is most useful in thinking about the appropriate specification of a time-series model (as opposed to the technical impediments to estimation discussed in the previous section). Thus, although the generalized least squares solution discussed above correctly deals with estimation problems, an impulse response analysis shows the potential defects of this model. In particular, a unit change in x has an immediate impact on y of β units, with that impact disappearing immediately. But a one-unit change in ε , the error, has an immediate one-unit impact on y , which decays at a rate of $(1 - \rho)$ percent per period. Thus, the error term has very complicated dynamics, but the variable of interest, x , is posited to have only an immediate impact, with no dynamics. Although this could be a correct description of the process, on its face it seems implausible. The lagged dependent variable

model, on the other hand, shows the same dynamics and decay rate for changes in both x and the error term; on its face, this is more plausible.

There are clearly many possible specifications that analysts must consider. These can be compared in terms of both their theoretical properties as well as how well they fit the data. In either enterprise, models with a lagged dependent variable will often be strong competitors. Thus, it is important to remember that models that contain a lagged dependent variable *and* dependent errors cannot even be estimated consistently by OLS (and so any generalized least squares procedures that start with OLS estimates will also be wrong). Fortunately, it is easy to use the Lagrange multiplier tests to examine whether the errors are correlated in models with lagged dependent variables; nicely, such tests work with the OLS residuals. Although such tests often show no remaining correlation of the errors, analysts must be rigorous in using them before continuing work.

NEW DIRECTIONS

The analysis of a single stationary time series is now well understood. It is often the case that OLS works well for such series, and a rigorous Lagrange multiplier testing strategy always indicates when OLS will not work well. When it does, time-series models are easy to estimate, and analysts can focus on finding the appropriate lag structure (sometimes theoretically, usually empirically) and studying the properties of the estimated model by impulse response and level change analyses. Theoretical interest rests on the choice of a particular lag structure to estimate.

Much research in the past decade has focused on nonstationary time series, particularly those with unit roots. Economists have concentrated on using these models to decompose time series into trends and cycles; such decompositions have long been of interest to macroeconomists. Attention has also focused on analyzing multiple time series, such as exchange rates between various countries, to see whether the trends and cycles they show are common or distinct. The estimation of all these models is technically difficult, but now implemented in standard time-series software packages. But for analysts trying to explain a single dependent variable that is nonstationary, the simple error correction model often suffices (and tests can indicate whether it does suffice). The advantage of

the error correction model is that it can be estimated by OLS.

Much attention has been paid recently to analyzing panel and time-series cross-section data. Panel data usually have relatively few repeated observations on the same individual, and so it is usually impossible to use sophisticated time-series models on the data. But attention must be paid to the interdependence of the observations on the same individual over time. Time-series cross-section data, however, typically have enough observations per unit so that it is possible to use at least some time-series methods and models to analyze the data. This is an area of great interest to those interested in comparative methods, whether in economics, sociology, or political science. Recently, there has been a spate of activity in the analysis of time-series cross-section data with a binary or other limited dependent variable; this activity has largely taken place in the study of political conflict. These are all areas of intense research, but much of that research begins with the analysis of a single time series.

—Nathaniel Beck

REFERENCES

- Beck, N., & Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, 89, 634–647.
- Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.
- Harvey, A. (1990). *The econometric analysis of time series* (2nd ed.). Cambridge: MIT Press.
- Mills, T. C. (1990). *Time series techniques for economists*. Cambridge, UK: Cambridge University Press.

TOBIT ANALYSIS

In defining Tobit, James Tobin (1958) quoted a passage about how nothing could not be less than nothing. Tobit permits analyzing dependent variables with zero as their lowest value and that are from CENSORED DATA, and it avoids violating ORDINARY LEAST SQUARES' (OLS) continuity and unboundedness assumptions about DEPENDENT VARIABLES. OLS can produce negative predicted values for such dependent variables and predict something that is less than nothing. Other consequences of using OLS

include biased coefficients and incorrectly obtaining statistically insignificant coefficients. Obtaining negative predicted values is the most common problem that Tobit solves, but it can analyze dependent variables with upper, upper and lower, and intermediate limits.

Tobit combines PROBIT ANALYSIS with OLS. Its coefficients allow identifying two effects on a dependent variable: (a) an effect on the probability of exceeding its limiting value(s), and (b) an effect on changing values beyond the limit (McDonald & Moffitt, 1980). For example, neighborhood crime rates cannot be less than zero, but they can have any positive value. For small areas, many will have crime rates of zero. OLS could predict negative crime rates, which are nonsensical (except for burglaries once a year). Manipulating Tobit coefficients permits identifying the effects of INDEPENDENT VARIABLES on the probability of a crime where the crime rate is zero and on changing the crime rate for places with nonzero rates. Tobit adjusts for the inability to measure a dependent variable's values beyond some limit, such as zero, and produces unbiased estimates of the independent variables' effects.

Tobit is a large-sample technique. The case-to-variable ratio should be 10 or more. It can be sensitive to the measurement of independent variables. These should have similar ranges even if transformations are necessary. Its sensitivity to heteroskedasticity can be tested and adjusted for by programs such as LIMDEP. The dependent variable should be continuous beyond its limit value, although Tobit can, at times, model more accurately the distributions of count variables than POISSON REGRESSION. Tobit requires the same signs for the effects it provides. The expected change in probability of having a crime in a no-crime area must be the same as that for the expected change in crime in areas with crimes. Otherwise, different techniques must be used.

The sum of the intercept and products of Tobit coefficients with their independent variables' values can be negative. Tobit coefficients are effects on an unobserved ("latent") variable. Their decomposition allows interpreting the effects on a dependent variable's observed values. The coefficients' sizes depend on the independent variables' measurement scales. Larger coefficients do not always indicate more important independent variables. Modifying Roncek's (1992) pseudostandardized coefficient by multiplying a Tobit coefficient only by the standard deviation of its independent variables adjusts for its measurement scale.

Dividing each coefficient by the dependent variable's standard deviation or the Tobit model's standard error is irrelevant.

—Dennis W. Roncek

REFERENCES

- McDonald, J. F., & Moffitt, R. A. (1980). The uses of Tobit analysis. *Review of Economics and Statistics*, 62, 318–321.
- Roncek, D. W. (1992). Learning more from Tobit coefficients. *American Sociological Review*, 57, 503–507.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24–36.

TOLERANCE

Tolerance is a measure that aids in assessing the degree of MULTICOLLINEARITY in a REGRESSION model. It indicates the amount of variation in an INDEPENDENT VARIABLE *not* statistically accounted for by the other independent variables in the model. If Tolerance = 1.0, then the independent variable in question has complete linear independence from the other independent variables in the model. If Tolerance = 0.0, then the independent variable in question has complete LINEAR DEPENDENCE on the independent variables in the model. The higher the Tolerance value, the less multicollinearity there is. Tolerance is routinely calculated in some statistical packages. In general, the analyst can regress each independent variable, j , on all the other independent variables in the model, not j , yielding an R -SQUARED accounting of X_j . Then, from this auxiliary regression, the $R^2_{X_j} = 1 - \text{Tolerance}$. For example, if $R^2_{X_j} = 0.65$, then Tolerance = 0.35.

—Michael S. Lewis-Beck

See also VARIANCE-INFLATION FACTORS

TOTAL SURVEY DESIGN

Total survey design is all of the elements of a social SURVEY that can affect estimates from that survey: the SAMPLE of people selected to be in the survey, the characteristics of those from whom data were

actually collected, the design of the questions, and the procedures that were used to ask questions and get answers.

The concept of total survey design pertains to social surveys that are designed to produce statistical descriptions of some POPULATION. Such surveys usually have the following characteristics: (a) A sample of the population to be described, not the whole population, is chosen to provide information; (b) only some fraction of those asked to provide information actually do so, the “RESPONDENTS”; (c) the data consist of the answers that chosen respondents provide to questions that are posed; and (d) the question-and-answer process is carried out either by having an interviewer ask questions and record answers or by having respondents themselves read questions, from a paper form or off of a computer, and record their answers on that paper form or directly into the computer. Each of those elements of the social survey has the potential to affect how well the results from the survey describe the population that is being studied. For example:

- a. Who has an actual chance to be selected to be in the sample, and how they are chosen, is critically important to how likely the sample is to mirror the population from which it is drawn.
- b. Those who actually respond to the questions may be systematically different from those who cannot or choose not to in ways that affect the survey results (see NONRESPONSE).
- c. The size of the sample, the number of people who actually respond, will affect the precision of estimates made from the sample results.
- d. Which questions are asked, and how well they are designed, will have a major effect on how well the answers to those questions provide valid information about what the survey is trying to measure.
- e. If interviewers are involved in data collection, the quality of INTERVIEWING—how well they carry out their roles—can affect the accuracy of the resulting data.
- f. If the questions are self-administered, with respondents reading them and providing answers themselves, how well the forms or programs are designed can affect the quality of data that result from the survey.

The fundamental premise of total survey design is that the data from a survey are no better than the worst elements of the methods used. When designing a survey, when carrying out a survey, or when evaluating the data from a survey, it is important to look at all of the elements of the survey that could affect data quality, the total survey design, when making choices about how to conduct that survey or estimating how accurate the resulting data are likely to be.

—Floyd J. Fowler, Jr.

REFERENCES

- Fowler, F. J. (2002). *Survey research methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

TRANSCRIPTION, TRANSCRIPT

Although it is now possible to store, code, and retrieve audio and audiovisual data in digital form, most interview-based QUALITATIVE RESEARCH continues to rely on the *transcription* of audiotaped material into textual form. Therefore, the production of accurate and high-quality *verbatim transcripts* is integral to establishing the credibility and trustworthiness (rigor) of qualitative research. Errors that reverse or significantly alter the meaning of what was said are the most problematic because they may lead the researcher to misinterpret and misquote respondents.

A review of sources of transcription errors (Poland, 1995) suggests that these arise most commonly (and are most difficult to detect) when alternative wordings, at least superficially, still make sense. This can occur, for example, when some words are mistaken for others with similar sound but very different meaning (e.g., “consultation” and “confrontation”); when inserting punctuation in ambiguous situations (e.g., “I hate it, you know. I do.” vs. “I hate it. You know I do.”); or when people mimic the speech of others (or conversations they have in their heads), and therefore, views that are not really their own appear as if they are. Transcriber fatigue; poor-quality recordings (due to background noise, use of low-quality tapes and microphones, poor microphone placement); and difficulty understanding accents and culturally specific turns of

phrase can also contribute to transcription errors. When material is produced by professional transcribers rather than by the researcher who conducted the interview, lack of familiarity with the subject matter or with the interview respondent and the interview itself often pose additional challenges.

Measures to enhance the quality of transcription include ensuring high-quality audio capture; attention to the selection and training of transcribers; the use of notation systems to capture pauses, emphasis, and other aspects of verbal interaction; MEMBER VALIDATION; and procedures for the post hoc review of transcription quality. The amount of time and energy put into producing verbatim transcripts (indeed, decisions about how verbatim is sufficient) depends on the requirements of the researcher. The needs of those engaged in CONVERSATION ANALYSIS, for example, typically exceed those of other qualitative researchers whose interest may be in the gist of what was said rather than in the analysis of the structure and form of naturally occurring talk.

Attention to issues of transcription quality need to be tempered with an understanding of what is inherently lost in the process of translating across media. Text on a page cannot recreate the full tenor and feel of an interview, and those who measure accuracy of transcription against the audiotapes from which they are derived would do well to remember that most nonverbal (and some verbal) communication is already missing from interview tapes. Indeed, a number of voices within the small but growing literature on transcription have sought to problematize the commonplace emphasis on the production of “verbatim” transcripts. Arguing for a more reflexive stance, they underscore the importance of seeing transcription as an interpretive act involving the selective emphasis and coding of data, and not simply as a technical issue of maximizing accuracy.

Finally, because of the inherent differences in standards of communicative competence between spoken and written language, an insistence on verbatim transcription that is carried over into the use of quotations in published research can inadvertently portray respondents as less articulate than they would be if they had written on the subject themselves. In cross-cultural and interclass research in particular, this can have important political and ethical ramifications.

—Blake D. Poland

REFERENCES

- Lapadat, J. C., & Lindsay, A. C. (1999). Transcription in research and practice: From standardization of technique to interpretive positionings. *Qualitative Inquiry*, 5(1), 64–86.
- Mishler, E. G. (1991). Representing discourse: The rhetoric of transcription. *Journal of Narrative and Life History*, 1(4), 255–280.
- Poland, B. D. (1995). Transcript quality as an aspect of rigor in qualitative research. *Qualitative Inquiry*, 1(3), 290–310.
- Poland, B. (2001). Transcription quality. In J. Gubrium & J. Holstein (Eds.), *Handbook of interview research*. Thousand Oaks, CA: Sage.

TRANSFER FUNCTION. See BOX-JENKINS MODELING

TRANSFORMATIONS

Transformations are used to change the characteristics of VARIABLES by reducing the number of categories, combining variables, or changing the metric of the variable. Prior to conducting the actual analysis of DATA, substantial amounts of time are spent in data processing. A major component of data processing involves variable transformation.

Transformations are used to collapse categories of nominal or ordinal data to obtain a smaller, more usable number of categories. If univariate statistics indicate that few observations fall in some categories of a variable, it may be sensible to combine them with similar categories. For nominal data, any categories can be combined, but for ordinal data, only adjacent categories should be combined. The decision to collapse categories is a compromise between having too many categories, with some containing few observations, and collapsing data into so few categories that valuable information is lost.

Transformations are also used to change the CODES on variables without collapsing categories. For example, log linear analyses are more easily interpreted if dichotomous categories are transformed into codes with 0 and 1. Categorical variables with three or more categories are often transformed to a series of DUMMY VARIABLES with only two outcomes (0 or 1) for use in MULTIVARIATE ANALYSES (Afifi & Clark, 1999). Transformations are also used to change the units of interval or ratio data, to sum or otherwise combine

variables to create a scale or other summary score, to create combinations of variables, and to summarize skip patterns.

One of the simplest operations is arithmetic transformation. All package programs used for data processing and analysis can perform addition (+), subtraction (−), multiplication (*), division (/), and exponentiation (**) (raising a variable to a power—for example, squaring it). In addition to the arithmetic operations, most software packages have numerous *functions* that can be used in data screening, transformations, and statistical analyses. Commonly used functions include taking the log of a variable; computing the mean of a variable; determining the maximum and minimum of a variable; assigning random numbers; recoding dates; and generating cumulative distribution functions of common DISTRIBUTIONS, such as the normal. These functions can be used in combination with the arithmetic operations. For example, “ $z = \log(y + 100)$ ” would result in a new variable called z being computed from the log of the sum of variable y plus 100.

Comparison operators are also heavily used in transforming variables or for case selection. These operators compare two or more variables, constants, or functions. They are used in *if-then* or *if-then-else* statements. Commonly used comparison operators include the following: gt , $>$ (greater than); ge , $\geq$ (greater than or equal to); lt , $<$ (less than); le , $\leq$ (less than or equal to); eq , $=$ (equal); ne , $\neq$ (not equal to); and, $\&$ (both or all true); or, $|$ (either true).

Finally, transformations are commonly used to achieve approximate univariate normality out of a SKEWED or nonnormal distribution. There are several strategies to determine the appropriate transformation. For example, the square root transformation is commonly used for data that follow a POISSON DISTRIBUTION. When the researcher does not know the theoretical distribution that the data follow, power transformations should be considered (Hoaglin, Mosteller, & Tukey, 1983; Tukey, 1977).

—Linda B. Bourque

REFERENCES

- Afifi, A. A., & Clark, V. A. (1996). *Computer-aided multivariate analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall.
- Bourque, L. B., & Clark, V. A. (1992). *Processing data: The survey example* (Sage University Paper series on

- Quantitative Applications in the Social Sciences, 07–085). Newbury Park, CA: Sage.
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

TRANSITION RATE

A transition rate, also known as a “transition intensity” and as an “instantaneous rate of transition,” is a concept that primarily arises when researchers want to distinguish COMPETING RISKS while conducting SURVIVAL ANALYSIS or EVENT HISTORY ANALYSIS. Interest in differentiating among competing risks originally came about in studies of the causes of death and failure. Not surprisingly, equivalent older names for a transition rate are the “cause-specific force of mortality (failure)” and the “cause-specific mortality (failure) rate.” Because the mathematical definition of a transition rate is closely related to the definition of a HAZARD RATE, some authors do not distinguish between the two terms.

Formally, a transition rate to a certain destination state k from a given origin state j at a particular time t is defined as

$$r_{jk}(t) = h_j(t)m_{jk}(t),$$

where $h_j(t)$ is the hazard rate of *any* transition from state j at time t and $m_{jk}(t)$ is the conditional probability that the transition (i.e., the event) consists of a change or move from state j to state k at time t . Like a hazard rate, a transition rate is measured as the inverse of time (e.g., per month or per year). Also like a hazard rate, a transition rate cannot be negative, but it may exceed one.

To give a few examples, if j denotes being married, k signifies being divorced, and time t refers to a person’s age in years, then $r_{jk}(t)$ would indicate the instantaneous rate per year at which a currently married person who is age t becomes divorced. As another example, if j denotes being unemployed, k represents having a job as a skilled worker, and t indicates the duration of a person’s unemployment in months, then $r_{jk}(t)$ would stand for the instantaneous rate per month at which a person who has been unemployed for a length of time t enters a job as a skilled worker.

ESTIMATION of a transition rate involves relatively minor modifications of the formulas for estimating the hazard rate of the occurrence of any event. A traditional demographic estimator of the transition rate associated with a specified discrete time interval $[u, v]$ is called the life table or actuarial ESTIMATOR; it is defined as

$$\widehat{r_{jk}(t)} = \frac{d_{jk}(u, v)}{(v - u)\{n_j(u) - \frac{1}{2}[d_{jk}(u, v) + c_j(u, v)]\}},$$

$$u \leq t < v,$$

where $d_{jk}(u, v)$ is the number of transitions from j to k in the interval $[u, v]$; $n_j(u)$ is the number that have survived in state j until time u ; and $c_j(u, v)$ is the number in state j at time u who are censored on the right before time v . The multiplier of one-half in the denominator on the right-hand side of this equation comes from assuming that the transitions and censored observations are spread uniformly over the interval $[u, v]$. A common, modern estimator is based on the Nelson-Aalen estimator of the integrated hazard rate:

$$\widehat{r_{jk}(t)} = \frac{d_{jk}(u, v)}{(v - u)n_j(u)}, \quad u \leq t < v.$$

For a mathematical introduction to these and other estimators of a transition rate, see Cox and Oakes (1984, Chap. 9). For a longer and less technical discussion addressed to social scientists, see Tuma and Hannan (1984, Chap. 3).

Typically, transition rates are modeled using a variant of a proportional hazard rate model (see Tuma & Hannan, 1984, Chaps. 5–8),

$$r_{jk}(t) = q_{jk}(t) \exp[\beta'_{jk}x_{jk}(t)],$$

where $q_{jk}(t)$ may be either a specific function of time (in a fully parametric model) or an unspecified nuisance function (in a partially parametric Cox model); $x_{jk}(t)$ is a possibly time-varying set of explanatory variables thought to affect the transition rate, and β_{jk} is a set of parameters giving the corresponding effects of the explanatory variables. Such models are usually estimated by MAXIMUM LIKELIHOOD or (in the case of the Cox model) by partial likelihood. Non-proportional models are often appropriate but are used less frequently in empirical applications; see Wu and Martinson (1993) for an example.

—Nancy Brandon Tuma

REFERENCES

- Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. New York: Chapman and Hall.
- Tuma, N. B., & Hannan, M. T. (1984). *Social dynamics: Models and methods*. Orlando, FL: Academic Press.
- Wu, L. L., & Martinson, B. C. (1993). Family structure and the risk of a premarital birth. *American Sociological Review*, 58(2), 210–232.

TREATMENT

This topic comes from the area of design and analysis of EXPERIMENTS. The term is used in two ways—either as an INDEPENDENT VARIABLE in the study of the effect of a variable on a DEPENDENT VARIABLE, or as a particular value of an independent variable. In spite of its double meaning, the term is so embedded in the literature on experiments that it will continue to be used. It is usually clear from the context whether the term is used as a variable or a value of a variable.

As a variable, the treatment is often a nominal-level variable. With just one experimental group and a control group, the units in the experimental group all get the treatment, whereas the units in the control group do not get the treatment. It is also possible to have more than two values of the independent variable. With two different teaching methods, the children in the first group get Treatment A, the children in the second group get Treatment B, and the children in the control group get no treatment. Here, treatment is used as a value of the NOMINAL VARIABLE. The outcome variable can be a reading score measured as a quantitative variable.

Data from an experiment of this type are commonly analyzed using ANALYSIS OF VARIANCE. These were the types of data with which Sir Ronald A. Fisher worked when he did his pioneering work on the design and analysis of agricultural experiments.

The treatment in an experiment can also be a QUANTITATIVE VARIABLE. For example, in a medical experiment, the treatment can be the injection of a medication. The first group can be patients given an injection of 5 mg, patients in the second group a dose of 10 mg, and so on. Again, the control group does not get any medication, except perhaps for an injection of sugar water as a placebo. The outcome variable can be a quantitative blood test of some type. Here, the term *treatment* has already been used in its second

meaning—as the particular value of an independent variable. A particular teaching method is a treatment, and a particular dosage of a medication is a treatment.

Data from an experiment of this latter type are commonly analyzed using REGRESSION analysis. A SCATTERPLOT will reveal whether the data fit a linear model or whether some other model is more appropriate.

The implication of the term *treatment* is that there is something in the research that can be manipulated by the researcher. Such manipulations are not used for observational data, and we do not see the term used for independent variables or values of these variables in sample surveys and other uses of observational data.

Apart from certain areas of psychology, experimentation is not anything commonly used in the social sciences, and the term *treatment* is therefore not used much in these sciences.

—Gudmund R. Iversen

REFERENCES

- Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters*. New York: Wiley.
- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Moore, D. S., & McCabe, G. P. (1999). *Introduction to the practice of statistics*. New York: W. H. Freeman.

TREE DIAGRAM

In graph theory, a *tree* is defined as a completely connected graph without cycles. Like any other graph, a tree diagram consists of a set of nodes connected by a set of arcs. Some nodes in a tree are internal nodes that have two or more arcs emanating from them. Other nodes are external nodes, or “leaves” of the tree, with only a single arc emanating from them. For example, in Figure 1(a), nodes *A*, *B*, *C*, and *D* are the leaves, and nodes *e*, *f*, and *g* are internal nodes. Sometimes, for graphical convenience, tree diagrams are presented in parallel form, as in Figure 1(b). Here, the arcs of the tree are represented only by the horizontal lines of the diagram, and the vertical lines are merely to connect the arcs.

In the social sciences, tree diagrams are used to represent relationships among concepts or entities. For example, tree diagrams may be used to represent the structure of certain social NETWORKS (Wasserman, 1994). Perhaps more commonly, trees are used to

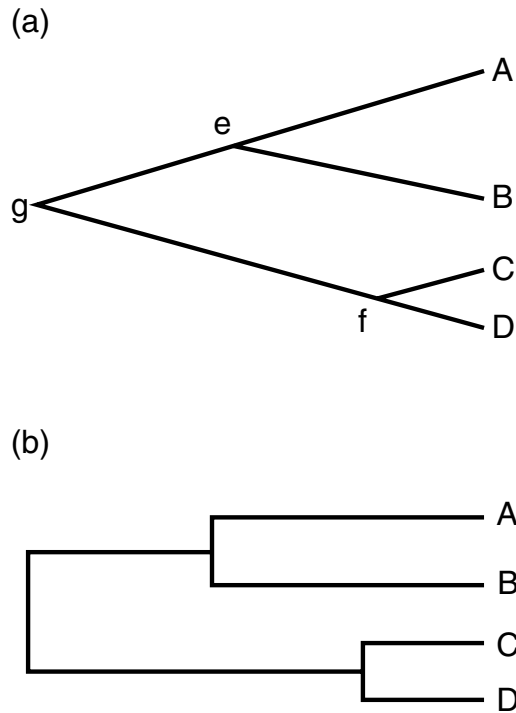


Figure 1 Two Displays of an Ultrametric Tree

represent the results of hierarchical CLUSTER ANALYSES. In these representations, usually the leaves of the tree represent the individual entities being clustered and the internal nodes of the tree graph represent sets of entities (i.e., clusters).

Tree diagrams are particularly useful for representing nested hierarchical relationships among entities. In the world, such hierarchical relationships may come about in a number of ways. This may come about when concepts or entities may be grouped at several different levels of generality. As a specific example, people may conceptually group kinship terms into clusters of terms describing nuclear family, other lineal relatives, and collateral relatives (Rosenberg & Kim, 1975). Another way hierarchical nested relationships can come about is when evolutionary-like processes of successive differentiation result in relationships among entities that can be represented by a branching tree structure, such as in the historical development of languages (e.g., Corter, 1996, Fig. 6.1).

Tree diagrams are sometimes used to model the proximity relationships among a set of N entities, such that distances in the tree (the model distances) approximate the proximities (the observed similarities or dissimilarities) between each pair of entities. Two

distinct types of tree structures have been used to model proximity relationships: the ultrametric tree and the additive tree (Corter, 1996).

The tree in Figure 1 is an ultrametric tree. Distances in an ultrametric tree satisfy the *ultrametric inequality*, which states that for any three objects x , y , and v ,

$$d_{x,y} \leq d_{x,v} = d_{y,v}$$

if objects x and y are neighbors in the tree (relative to v). Graphically, the leaves in an ultrametric tree are all equidistant from a special point in the graph, termed the *root* of the tree. In Figure 1(a), node g is the root of the tree.

The additive tree [Figure 2(a)] is a more general structure. Distances in an additive tree satisfy the *additive inequality*, which states that for any four objects x , y , u , and v ,

$$d_{x,y} + d_{u,v} \leq d_{x,v} + d_{y,u} = d_{x,u} + d_{y,v}$$

if (x, y) and (u, v) are relative neighbors in the tree structure. Graphically, the leaves in an additive tree may be of unequal lengths. This allows the additive tree to represent entities that differ in typicality. For

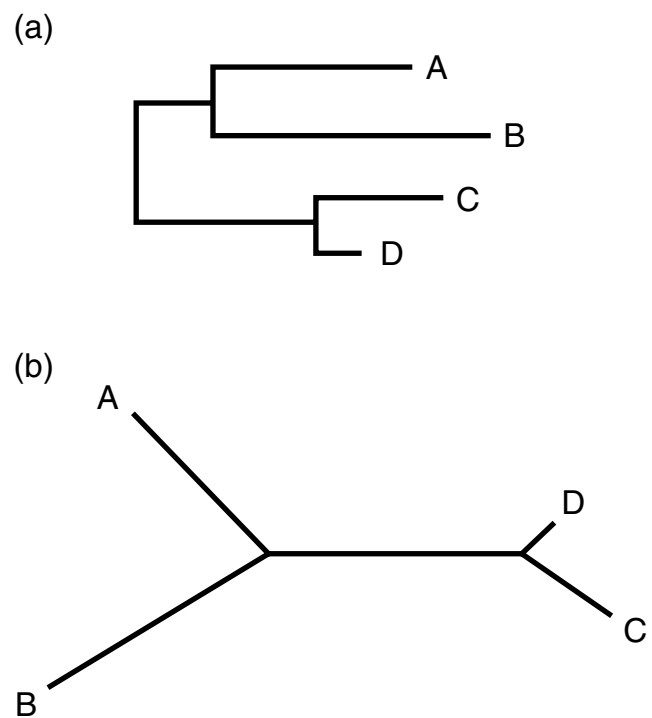


Figure 2 Two Displays of an Additive Tree

example, in Figure 2, objects *B* and *C* are relatively atypical, as shown by their relatively long leaf arcs. Additive trees are sometimes represented in unrooted form, as in Figure 2(b).

—James E. Corter

REFERENCES

- Corter, J. E. (1996). *Tree models of similarity and association* (Sage University Paper series on Quantitative Applications in the Social Sciences, 07–112). Thousand Oaks, CA: Sage.
- Rosenberg, S., & Kim, M. P. (1975). The method of sorting as a data-gathering method in multivariate research. *Multivariate Behavioral Research*, 10, 489–502.
- Wasserman, S. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

TREND ANALYSIS

A prominent feature of many TIME SERIES variables is that they tend to move upward or downward over time. This tendency may be defined as a trend. Two models frequently are employed to characterize trending data. The first model is a *deterministic trend*, $Y_t = \lambda_0 + \lambda_1 T + \varepsilon_t$, where Y_t is a time series, T is time, λ_0 and λ_1 are PARAMETERS, and ε_t is a STOCHASTIC shock $\sim N(0, \sigma^2)$. If this model generates Y_t , the residuals from a REGRESSION of Y_t on T constitute a trend-stationary process.

The second model is a *stochastic trend*, $Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$, where Y_{t-1} is the time series lagged one period, β_0 and β_1 are parameters, and ε_t is a stochastic shock. If $\beta_0 = 0$, and $\beta_1 = 1$, this model is a RANDOM WALK; otherwise, it is a random walk with drift. Recursive substitution shows that a random walk equals an initial value Y_0 , plus the sum of all shocks, $\sum \varepsilon_i$. The random walk with drift also includes $\beta_0 T$, which imparts a deterministic component to the process. First-differencing (i.e., $\Delta Y_t = Y_t - Y_{t-1}$) a random walk or random walk with drift produces a difference-stationary process.

Trending variables are *nonstationary* because they lack constant means, variances, and covariances. Regressing one nonstationary variable on another poses threats to inference (“spurious regressions”). Thus, it is vital to detect trends (nonstationarity) in time-series data. A trending series has a “characteristic

signature” when correlated with lags of itself. Several such correlations constitute an *AUTOCORRELATION function* (ACF). The ACF for a trending series has large correlations at low lags, and correlations decline slowly at longer lags. Formal statistical tests (UNIT-ROOT tests) also are widely used to detect nonstationarity.

If nonstationarity is suspected, conventional practice is to difference the data (assuming stochastic, not deterministic, trends). To determine if variables move together in the long run, COINTEGRATION tests are performed. Cointegrating variables can be modeled in ERROR correction form—for example, $\Delta Y_t = \beta_0 + \beta_1 \Delta X_t - \alpha(Y_{t-1} - \lambda X_{t-1}) + \varepsilon_t$ —where ΔY_t is first-differenced Y ; ΔX_t is first-differenced X ; $(Y_{t-1} - \lambda X_{t-1})$ is an error correction mechanism for the long-term relationship between Y and X ; ε_t is a stochastic shock; and β_0 , β_1 , α , and λ are parameters. α measures the speed with which shocks to the system are eroded by the cointegrating relationship between Y and X . α should be negatively signed and less than 1.0 in absolute value.

Differencing to ensure stationarity has been criticized on theoretical and statistical grounds. The claim is that many social science time series are either “near integrated” (β_1 in a random walk model is slightly less 1.0) or “fractionally integrated.” A fractionally integrated series possesses “long memory” and may be nonstationary. Because unit-root tests have low statistical power, they may not detect near or fractionally integrated processes. Estimating fractional differencing parameter (d) in rival ARFIMA (autoregressive, fractionally integrated, moving average) models is a useful way of considering the possibility that near or fractionally integrated processes may be operative. Recent generalizations of the concepts of cointegration and error correction permit one to model fractionally cointegrated processes.

—Harold D. Clarke

REFERENCES

- Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, 73, 5–59.
- DeBoef, S., & Granato, J. (2000). Testing for cointegrating relationships with near-integrated data. *Political Analysis*, 8, 99–117.
- Franses, P. H. (1998). *Time series models for business and economic forecasting*. Cambridge, UK: Cambridge University Press.

Hendry, D. F. (1995). *Dynamic econometrics*. Oxford, UK: Oxford University Press.

TRIANGULATION

Triangulation refers to the use of more than one approach to the investigation of a research question in order to enhance confidence in the ensuing findings. Because much social research is founded on the use of a single research method, and as such may suffer from limitations associated with that method or from the specific application of it, triangulation offers the prospect of enhanced confidence. Triangulation is one of the several rationales for multimethod research. The term derives from surveying, where it refers to the use of a series of triangles to map out an area.

TRIANGULATION AND MEASUREMENT

The idea of triangulation is very much associated with measurement practices in social and behavioral research. An early reference to triangulation was in relation to the idea of UNOBTRUSIVE METHOD proposed by Webb, Campbell, Schwartz, and Sechrest (1966), who suggested, "Once a proposition has been confirmed by two or more independent measurement processes, the uncertainty of its interpretation is greatly reduced. The most persuasive evidence comes through a triangulation of measurement processes" (p. 3). Thus, if we devise a new survey-based measure of a concept such as emotional labor, our confidence in that measure will be greater if we can confirm the distribution and correlates of emotional labor through the use of another method, such as structured observation. Of course, the prospect is raised that the two sets of findings may be inconsistent, but as Webb et al. observed, such an occurrence underlines the problem of relying on just one measure or method. Equally, the failure for two sets of results to converge may prompt new lines of inquiry relating to either the methods concerned or the substantive area involved. A related point is that even though a triangulation exercise may yield convergent findings, we should be wary of concluding that this means that the findings are unquestionable. It may be that both sets of data are flawed.

TYPES OF TRIANGULATION

Denzin (1970) extended the idea of triangulation beyond its conventional association with research methods and designs. He distinguished four forms of triangulation:

1. *Data triangulation*, which entails gathering data through several sampling strategies so that slices of data at different times and in different social situations, as well as on a variety of people, are gathered
2. *Investigator triangulation*, which refers to the use of more than one researcher in the field to gather and interpret data
3. *Theoretical triangulation*, which refers to the use of more than one theoretical position in interpreting data
4. *Methodological triangulation*, which refers to the use of more than one method for gathering data

The fourth of these, as the preceding discussion implies, is the most common of the meanings of the term. Denzin drew a distinction between *within-method* and *between-method* triangulation. The former involves the use of varieties of the same method to investigate a research issue; for example, a self-completion questionnaire might contain two contrasting scales to measure emotional labor. Between-method triangulation involved contrasting research methods, such as a questionnaire and observation. Sometimes, this meaning of triangulation is taken to include the combined use of QUANTITATIVE RESEARCH and QUALITATIVE RESEARCH to determine how far they arrive at convergent findings. For example, a study in the United Kingdom by Hughes et al. (1997) of the consumption of "designer drinks" by young people employed both structured interviews and focus groups. The two sets of data were mutually confirming in that they showed a clear pattern of age differences in attitudes toward these types of alcoholic drinks.

Triangulation is sometimes used to refer to all instances in which two or more research methods are employed. Thus, it might be used to refer to multimethod research, in which a quantitative and a qualitative research method are combined to provide a more complete set of findings than could be arrived at through the administration of one of the methods alone. However, it can be argued that there are good reasons

for reserving the term for those specific occasions in which researchers seek to check the VALIDITY of their findings by cross-checking them with another method.

EVALUATION OF TRIANGULATION

The idea of triangulation has been criticized on several grounds. First, it is sometimes accused of subscribing to a naive REALISM that implies that there can be a single definitive account of the social world. Such realist positions have come under attack from writers aligned with CONSTRUCTIONISM and who argue that research findings should be seen as just one among many possible renditions of social life. On the other hand, writers working within a constructionist framework do not deny the potential of triangulation; instead, they depict its utility in terms of adding a sense of richness and complexity to an inquiry. As such, triangulation becomes a device for enhancing the credibility and persuasiveness of a research account. A second criticism is that triangulation assumes that sets of data deriving from different research methods can be unambiguously compared and regarded as equivalent in terms of their capacity to address a research question. Such a view fails to take account of the different social circumstances associated with the administration of different research methods, especially those associated with a between-methods approach (following Denzin's, 1970, distinction). For example, the apparent failure of findings deriving from the administration of a STRUCTURED INTERVIEW to converge with FOCUS GROUP data may have more to do with the possibility that the former taps private views as opposed to the more general ones that might be voiced in the more public arena of the focus group.

Triangulation has come to assume a variety of meanings, although the association with the combined use of two or more research methods within a strategy of convergent validity is the most common. In recent years, it has attracted some criticism for its apparent subscription to a naively realist position.

—Alan Bryman

REFERENCES

- Denzin, N. K. (1970). *The research act in sociology*. Chicago: Aldine.
- Hughes, K., MacKintosh, A. M., Hastings, G., Wheeler, C., Watson, J., & Inglis, J. (1997). Young people, alcohol, and designer drinks: A quantitative and qualitative study. *British Medical Journal*, 314, 414–418.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive measures in the social sciences*. Chicago: Rand McNally.

TRIMMING

A fundamental problem with the sample mean is that slight departures from normality can result in relatively poor power (e.g., Staudte & Sheather, 1990; Wilcox, 2003), a result that follows from a seminal paper by Tukey (1960). Roughly, the sample variance can be greatly inflated by even one unusually large or small value (called an OUTLIER), which in turn can result in low power when using means versus other measures of central tendency that might be used.

There are two general strategies for dealing with this problem, and each has advantages and disadvantages. The first strategy is to check for outliers, discard any that are found, and use the remaining data to compute a measure of location. The second strategy is to discard a fixed proportion of the largest and smallest observations and then average the rest. This latter strategy is called a trimmed mean, and, unlike the first strategy, empirical checks on whether there are outliers are not made.

A 10% trimmed mean is computed by removing the smallest 10% of the observations, as well as the largest 10%, and averaging the values that remain. To compute a 20% trimmed mean, remove the smallest 20%, the largest 20%, and compute the usual sample mean using the data that remain. The sample mean and the usual median are special cases that represent two extremes: no trimming and the maximum amount.

A basic issue is deciding how much to trim. In terms of achieving a low standard error relative to other measures of central tendency that might be used, Rosenberger and Gasko (1983) recommend 20% trimming, except when sample sizes are small, in which case they recommend 25% instead. This recommendation stems in part from the goal of having a relatively small standard error when sampling from a normal distribution. The median, for example, has a relatively high standard error under normality, which in turn can mean relatively low power when testing hypotheses. This recommendation is counterintuitive based on standard training, but a simple explanation can be

found in Wilcox (2001, 2003). Moreover, theory and simulations indicate that as the amount of trimming increases, known problems with means, when testing hypotheses, diminish. This is not to say that 20% is always optimal; it is not, but it is a reasonable choice for general use.

After observations are trimmed, special methods for testing hypotheses must be used (and the same is true when discarding outliers). That is, applying standard methods for means to the data that remain violates basic principles used to derive these techniques (see Wilcox, 2003).

The notion of trimming, to avoid problems with outliers and nonnormality, has been used in regression as well. For comments on the relative merits of this approach, versus other robust regression estimators that have been proposed, see Wilcox (2003).

—Rand R. Wilcox

REFERENCES

- Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis*. New York: Wiley.
- Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.
- Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.
- Wilcox, R. R. (2001). *Fundamentals of modern statistical methods*. New York: Springer.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.

TRUE SCORE

True score is assumed in measurement theory, which postulates that every measurement X consists of two parts: the respondent's true value or true score, and random error. That is, $X = T + e$. This concept relates to RELIABILITY. A measure with no random errors corresponds perfectly with its true score and has perfect reliability. Conversely, a measure with all random error does not correspond to any true score and has zero reliability.

—Tim Futing Liao

TRUNCATION. See CENSORING AND TRUNCATION

TRUSTWORTHINESS CRITERIA

Trustworthiness criteria provide yardsticks by which the rigor, VALIDITY, systematic nature of methods employed, and findings of qualitative studies may be judged. It is the name given to criteria appropriate to emergent, interpretive, nonpositivist, nonexperimental models of inquiry that frequently are labeled constructivist, critical theorist, action inquiry, phenomenological CASE STUDIES, or naturalistic paradigms. These criteria were the first to be proposed (Guba, 1981; Lincoln & Guba, 1985) in response to serious queries regarding the soundness, RELIABILITY, and rigor of qualitative research done under new-paradigm models.

Conventional (positivist) criteria of rigor are well established and understood: INTERNAL VALIDITY, EXTERNAL VALIDITY, reliability (replicability), and OBJECTIVITY. These four rigor domains respond, respectively, to an isomorphism to reality, to generalizability, to the stability of findings (and therefore, the ability of any given study to be replicated), and to freedom from investigator bias. Phenomenological, or new-paradigm, inquiry, however, specifically rejects each of those four domains as unachievable (in either a practical or a philosophical sense) or inappropriate (in a paradigmatic sense). It has proposed in their stead four criteria that are more responsive to both the axiomatic requirements of phenomenological inquiry and the requirements of methodologies and design strategies often more appropriate to such research, evaluation, or policy analyses, which are frequently, although not always, qualitative in nature.

Early efforts to create standards for judging non-quantitative and emergent-paradigm inquiries elicited the trustworthiness criteria: *credibility*, *transferability*, *dependability*, and *confirmability*. These criteria, deemed “trustworthiness criteria” to signal that attention to them increased confidence in the rigorousness of findings, were devised as parallel to the four criteria of conventional research. These criteria, designed as parallel to criteria for experimental models, have been criticized as “foundational,” that is, tied to a foundation emanating from conventional inquiry rather

than from characteristics inherent in phenomenological (constructivist) inquiry. Like conventional criteria for rigor, these criteria reference primarily methodological concerns; because of that, they are still highly useful when examining the methodological adequacy of case studies.

Credibility refers to the plausibility of an account (a case study): Is it believable? Is the report deemed an accurate statement by those whose LIVED EXPERIENCE is reported? *Transferability* refers to the report's ability to be utilized in (transferred to) another setting similar to that in which the original case was conducted. Judgments about transferability are unlike those about GENERALIZABILITY, however, in one important aspect. In conventional research, generalizability estimates are asserted by the researchers. In assessing transferability, however, judgments about the usefulness of results are made by "receiving contexts" (Guba & Lincoln, 1989; Lincoln & Guba, 1985), such as users and consumers of the research findings, and individuals who wish to utilize the findings in their own situations. *Dependability* references the stability, trackability, and logic of the research process employed. Dependability is assessed through a process audit, which analyzes the methodological decisions made and certifies their soundness. *Confirmability* certifies that the data reported in the case can be pursued all the way back to original data sources (e.g., FIELDNOTES, documents, records, observation logs, etc.).

Trustworthiness criteria are primarily methodological criteria; they are responsive to method (although some of the suggested methods—Guba & Lincoln, 1989—are directly tied to phenomenological premises, e.g., member checking to ascertain whether the researcher has understood the respondent's social constructions), rather than directly tied to paradigmatic assumptions. However, because new-paradigm inquiry often demands nonconventional, qualitative methods, and because trustworthiness criteria are tied heavily to the use of such methods, they continue to have applicability in judging the rigor (for instance, the level of engagement and persistence in the content) of an essential element—method—of emergent-paradigm inquiry.

—Yvonna S. Lincoln

REFERENCES

- Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communications and Technology Journal*, 29, 75–92.
- Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

T-RATIO. See T-TEST

T-STATISTIC. See T-DISTRIBUTION

T-TEST

It was William Gosset who, when performing quality controls at the Guinness brewery in Dublin, wrote articles in mathematical statistics under the pseudonym "Student." His 1908 paper "The Probable Error of a Mean" provided the basis of Student's *t*-test. The use of his Student's *t*-DISTRIBUTION in many applications is an important contribution to the analysis of small samples in mathematical statistics.

STUDENT'S T-TEST OF A SINGLE MEAN

A simple example of such an application is the test of a mean. When σ is unknown and sample size is large, we can be confident that the estimated standard error of the mean $\hat{\sigma}_{\bar{X}}$ (standard deviation of the sampling distribution of means under H_0) will be equal to the true $\sigma_{\bar{X}}$, and the ratio we will evaluate and that will be referred to as a normal sampling distribution is the standardized score $z = [\bar{X} - E(\bar{X})]/\sigma_{\bar{X}}$. However, when sample size is small, we cannot have this confidence in our estimate of the standard error; then, the ratio we use is not a *z*-score, but $t = [\bar{X} - E(\bar{X})]/\hat{\sigma}_{\bar{X}}$, in which the denominator is not a constant but a random variable, which varies along with the sample size, that is, with the DEGREES OF FREEDOM.

The *t*-distribution has a mean of 0 and a symmetric and unimodal shape, but it is flatter in the middle and denser at the extremes in comparison with the NORMAL DISTRIBUTION. As sample size *n* grows large, the distribution of *t* approaches the standardized normal distribution.

Now, suppose we hypothesize that the mean income of a particular group, say, the chemists, is larger than 2,000 euros per month ($H_a: \mu > 2,000$) as against the null hypothesis $H_0: \mu = 2,000$. The standard deviation σ is unknown. We draw a small RANDOM SAMPLE of $n = 26$. We obtain $\bar{X} = 2,040$ and $s = 100$. Our estimate of $\hat{\sigma}_{\bar{X}}$ will be

$$\frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n-1}},$$

and consequently, the t -ratio is

$$t = \frac{2,040 - 2,000}{100/\sqrt{25}} = 2.$$

In the t -table, we find for a one-tailed test and $\alpha = 0.05$ a critical t -value of $t^* = 1.706$. Our sample value is higher, hence, the mean income of chemists is significantly higher than 2,000. Even in such a small sample, we obtain a significant result. However, had we performed a two-tailed test ($t^* = 2.056$), then the result would have been nonsignificant.

STUDENT'S T-TEST OF DIFFERENCE OF MEANS

Another example of application of Student's t is the t -test of DIFFERENCE OF MEANS. We can think of two groups in an experimental design, such as a group of men and a group of women, for which the scores of a QUANTITATIVE VARIABLE are considered, let us say their smoking behavior measured as the number of cigarettes per day. In this example, the dependent variable Y (smoking) is CONTINUOUS, and the independent variable X (representing the two groups) is DICHOTOMOUS. We could give the experimental group and the control group the codes $X = 1$ and $X = 2$, respectively. We want to investigate whether the mean number of cigarettes for men ($X = 1$) is significantly different from (two-tailed test) or significantly lower than (one-tailed test) the mean number of cigarettes for women ($X = 2$). The null hypothesis states that the difference of means is equal to zero. Take two groups of five individuals, that is, $n_1 = n_2 = 5$. Let the series of scores for the total sample be 2 2 3 4 4 5 6 7 8 9 with mean 5 and variation 54. The group of men with scores 2 2 3 4 4 has a mean of 3 and a variation of 4. The group of women with scores 5 6 7 8 9 has a mean of 7 and a variation of 10. The dispersions are unequal: variations 4 and 10. In order to perform a t -test legitimately, the

distributions should be approximately normal (in the case of small groups), and the variances must not be significantly different. For both conditions, there exist a couple of test procedures, the test of normality and the test of homoskedasticity (LEVENE'S TEST), respectively. Assuming that both conditions are fulfilled, the calculations are as follows.

A mean (pooled average) of the two variances is calculated:

$$\begin{aligned} s_w^2 &= \frac{\text{variation group 1} + \text{variation group 2}}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{4 + 10}{(5 - 1) + (5 - 1)} = 1.75. \end{aligned}$$

In order to obtain the sampling distribution, we consider all theoretically possible pairs of random and independent samples (with sizes n_1 and n_2) that could have been drawn from a theoretical population (i.e., a population under H_0 for which the difference in means of both samples would be zero or, at most, purely random). For each of the countless number of pairs of samples, we consider the difference of mean number of cigarettes. In this way, we obtain a sampling distribution of differences of means. The variance of such a sampling distribution appears to be equal to the sum of the variances of the two sampling distributions of means separately (one for samples with size n_1 and one for samples with size n_2). Thus,

$$\sigma_{\text{diff}}^2 = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}.$$

The term σ^2 , that is, the variance of the theoretical population out of which both samples would have been drawn according to H_0 , is not known, and it is therefore estimated. The estimator used is the pooled average s_w^2 , which was calculated above. In order to use statistical tables, we apply standardization: From each difference of means, the mean of the sampling distribution of differences of means (which is equal to 0 under H_0) is subtracted, and the result is divided by the standard error σ_{diff} (which is equal to the square root of the estimated variance of the sampling distribution). Instead of classical z -scores, we obtain t -scores; the reasons for this are that the size of the sample is small, and that the variance of the population must be estimated, which results in a deviation from the normal distribution. If we now calculate the t -score for our pair of samples,

we can then situate it in the sampling distribution and come to a conclusion:

$$t = \frac{|\bar{Y}_{(1)} - \bar{Y}_{(2)}| - 0}{\left(\frac{s_w^2}{n_1} + \frac{s_w^2}{n_2}\right)^{1/2}} = \frac{|3 - 7| - 0}{\left(\frac{1.75}{5} + \frac{1.75}{5}\right)^{1/2}} = 4.78.$$

For an a priori postulated Type I error of $\alpha = 0.05$ and for $(n_1 - 1) + (n_2 - 1) = (5 - 1) + (5 - 1) = 8$ degrees of freedom, and using a one-tailed test (i.e., supposing that women smoke more than men nowadays), we find in the t -table a critical value of $t^* = 1.86$. The absolute value of t that we have calculated is higher and, consequently, has a lower than 5% probability of occurrence under the null hypothesis, that is, it is situated in the rejection region. Thus, with an a priori 95% likelihood of not being mistaken, there is a significant difference between the mean smoking behavior of women and men, with women smoking more on the average.

SOME CONSIDERATIONS

The example used here for the difference of means test is the same as in the entry on the F ratio, and the reader can verify that the F value of 22.86 obtained there is (within rounding errors) equal to the square of the t -value found of 4.78, because when the degree of freedom is one, $F = t^2$.

The null hypothesis for the test of difference between two means stated that this difference is zero. It is also possible to use some other number than zero and to hypothesize that the difference of means differs from that number.

When more than two samples are investigated, we are not allowed to make all the group comparisons by means of several t -tests; instead, we have to perform one F test. If we really want to perform multiple comparisons known as tests of contrasts, then we have to test more conservatively, that is, with lower probability of TYPE I ERROR (see ANALYSIS OF VARIANCE and related topics).

The reader should not forget that the distribution of t can be used only if the requirement that the population distribution is normal is fulfilled. This holds for the test of a single mean as well as for the test of difference of means. In most cases, we have no information on the population distribution, so this normality assumption has to be tested in the sample.

In our example, we assumed homogeneity of variances. However, this assumption can be tested by means of an F test. If this test results in a significant

difference between the variances s_1^2 and s_2^2 , then the procedure of calculating a pooled average s_w^2 of the two variances, which was used as an estimator of σ^2 , has to be replaced by a procedure in which separate standard errors are computed from each sample. The appropriate number of degrees of freedom then has to be corrected. The formula for this can be found in some statistical textbooks, such as Hays (1972) and Blalock (1972).

It should be mentioned that the two samples in our example have to be independent samples. A different procedure must be chosen when observations are paired, such as when individuals are matched in pairs by the researcher or, to take our example, when some men in Group 1 and some women in Group 2 are husband and wife. In such a case, our statement that

$$\sigma_{\text{diff}}^2 = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$$

no longer holds, but a covariance will be involved in the calculation of σ_{diff}^2 . The procedure to follow here is to look at the data as coming from one sample of pairs and to proceed with the t -test for a single mean.

A final remark is that the t -test is used in more advanced analysis, such as MULTIPLE REGRESSION for significance testing of parameters. Here, too, this t -test can be replaced by an F test, where F will be the square of t . We can also mention that Hotelling has developed a more extended version of t , Hotelling's T^2 test, which is used in DISCRIMINANT ANALYSIS for two groups and many variables.

—Jacques Tacq

REFERENCES

- Blalock, H. M. (1972). *Social statistics*. Tokyo: McGraw-Hill Kogakusha.
- Hays, W. L. (1972). *Statistics for the social sciences*. New York: Holt.
- Student. (1908). *The probable error of a mean*. In E. S. Pearson & J. Wishart (Eds.), *Student's collected papers*. London: University College.
- Tacq, J. J. A. (1997). *Multivariate analysis techniques in social science research: From problem to analysis*. London: Sage.

TWENTY STATEMENTS TEST

Self and identity are two concerns of both psychology and sociology, enabling the researcher to see how people place themselves in social categories and

reflect upon who they are. The Twenty Statements Test (sometimes called the “Who am I?” test) is one of many instruments devised to measure the concept of self. It was introduced by symbolic interactionist sociologists Manford Kuhn and Thomas McPartland (1954), who wished to make George Herbert Mead’s (a leading pragmatic philosopher) concept of the self more rigorous and operational.

Subjects are asked to write down the first 20 statements that come into their mind in answer to the question “Who Am I?” They are told that there are no right or wrong answers. The test is an open-ended instrument in which answers listed can be nouns or adjectives. In their early work, Kuhn and McPartland predicted that with increasing age and with growing social change, adjectival responses (e.g., cheerful, religious, existentially aware) would be used more frequently than substantive noun responses (e.g., wife, student, man).

Theorists of self are especially interested in how ideas about self change. In traditional worlds, “pre-modern identities” may often be taken for granted. People know who they are, defining themselves through traditions and religion, with identities being shared and communal, given as part of the world. Such identities are not seen as a problem and are usually just taken for granted. By contrast, as modernity develops, human individuality and individual identity become more important. Increasingly, self-reflection upon just who one is becomes an issue. The Twenty Statements Test was meant as a research tool into such questions.

The instrument is one of a great many tools for measuring the self and is not used very much these days. Although it does allow for a broad, sweeping depiction of a sense of self, it is very basic. It has been criticized for not taking into account (a) the situated nature of the self, which changes from context to context; (b) the role of significant others in shaping what can be said; and (c) the problems of coding and content analysis, which are problems commonly associated with an open-ended questionnaire.

—Ken Plummer

See also SELF-REPORT MEASURE, SYMBOLIC INTERACTIONISM

REFERENCES

- Bracken, B. (Ed.). (1995). *Handbook of self concept: Developmental, social and clinical consequences*. New York: Wiley.
- Hewitt, J. (1989). *Dilemmas of the American self*. Philadelphia: Temple University Press.
- Kuhn, M., & McPartland, T. S. (1954). An empirical investigation of self attitudes. *American Sociological Review*, 19, 68–76.
- Wylie, R. (1974). *The self concept: A review of methodological considerations and measuring instruments*. Lincoln: University of Nebraska Press.

TWO-SIDED TEST. See TWO-TAILED TEST

TWO-STAGE LEAST SQUARES

Two-stage least squares is a procedure commonly used to estimate the PARAMETERS of a SIMULTANEOUS EQUATION model. In the face of reciprocal causation, a given in the general simultaneous equation model, where variables influence each other, ORDINARY LEAST SQUARES will yield BIASED estimates. However, two-stage least squares (TLS, or 2SLS), an INSTRUMENTAL VARIABLES method, will yield consistent parameter ESTIMATION in this situation. Suppose the following model:

$$Y_1 = b_0 + b_1 X_1 + b_2 Y_2 + e, \quad (1)$$

$$Y_2 = a_0 + a_1 Y_1 + a_2 X_2 + u. \quad (2)$$

TLS estimation goes forward in stages, and it estimates parameters an equation at a time. Let us estimate equation (1). In the first stage, all of the EXOGENOUS VARIABLES in the system (X_1 , X_2) are regressed on the independent ENDOGENOUS VARIABLE (Y_2). This creates an instrumental variable, predicted- Y_2 , which is strictly a function of the exogenous variables. In the second stage, predicted- Y_2 is substituted for observed- Y_2 in equation (1), and the equation is estimated with ordinary least squares. These second-stage estimates are consistent, because all of the INDEPENDENT VARIABLES in the equation are effectively exogenous. The same two-stage procedure is then applied to equation (2) to obtain estimates for the parameters in that equation.

—Michael S. Lewis-Beck

See also CAUSAL MODELING, INSTRUMENTAL VARIABLE

TWO-TAILED TEST

A NULL HYPOTHESIS most often specifies one particular value of a population PARAMETER, even though we do not necessarily believe this value to be the true value of the parameter. In the study of the relationship between two variables, using REGRESSION analysis as an example, the null hypothesis used most often is that there is no relationship between the two variables. That translates into the null hypothesis $H_0 : \beta = 0$, where β is the SLOPE of the population regression line. Most often, we study the RELATIONSHIP between two variables for the purpose of demonstrating that a relationship does exist between the two variables. This is achieved if we are able to reject the null hypothesis of no relationship.

If we can show that the slope of the population regression line is not equal to zero—by rejecting the null hypothesis—then does the slope have a positive or negative value? If we do not have any extra information about the slope, then the alternative hypothesis becomes $H_a : \beta \neq 0$, meaning the slope can be negative or positive. Such an alternative hypothesis makes the test a *two-tailed* test, also known as a *two-sided* test.

In the original approach to this problem, as developed by Sir Ronald A. Fisher and others, they chose a preset SIGNIFICANCE LEVEL, say $\alpha = 0.05$, for lack of any better value. Depending on the nature of the analysis, they chose a test statistic, usually a z , t , CHI-SQUARE or F . For the normal z variable, 2.5% of the values are less than -1.96 and 2.5% of the values are larger than 1.96 . Thus, for the normal z variable, the null hypothesis would be rejected if the observed value of z was less than -1.96 or larger than 1.96 , meaning that the value fell in one of the two tails of the distribution.

This means that for a true null hypothesis, and repeating the study a large number of times, we would get samples, and thereby z s, 5% of the time either less than -1.96 or larger than 1.96 . For these samples, we would erroneously reject the null hypothesis. The reason for the name “two-tailed test” is that the null hypothesis is rejected for unusually large or small values of the test statistic z or t . Tests using the chi-square or the F statistic are, by their nature, set up to be two-tailed tests.

With the arrival of statistical software for the analysis of the data, there has been a shift away from the significance level α , chosen before the analysis, to

the use of P VALUES, computed as part of the analysis. For z and t , if we see the absolute value symbol $|z|$ or $|t|$, then the corresponding p values represent two-tailed tests. If we decided before starting the analysis to make use of a two-tailed test, and the software provides a one-tailed p value, then we have to double this p value.

It is also possible to simply report the computed p value and let the reader determine whether the test should be two-tailed or one-tailed.

—Gudmund R. Iversen

REFERENCES

- Mohr, L. B. (1990). *Understanding significance testing* (Sage University Paper series on Quantitative Applications in the Social Sciences, 07-073). Newbury Park, CA: Sage.
- Morrison, D. E., & Henkel, R. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.

TWO-WAY ANOVA

Two-way analysis of variance (two-way ANOVA) is the test used to analyze the DATA from a study in which the investigator wishes to examine both the separate and the combined effects of two VARIABLES on some measure of behavior. The data come from a factorial experiment or study in which separate levels of each of two variables or factors (e.g., Factor A and Factor B) are identified and all combinations are formed. Factors A and B may be manipulated INDEPENDENT VARIABLES, such as incentive level or performing alone versus in groups, or measured levels of subject variables, such as age and sex. Among the more interesting applications in the social sciences are ones in which Factor A is a manipulated independent variable and Factor B is a subject variable. The primary interest is then whether the effect of Factor A on the response measure differs for individuals defined by different levels of Factor B.

Consider, for example, a study of the effects of exposure to violent video games on subsequent aggressive behavior exhibited by young boys and girls. Let's say that the behavior being recorded—the DEPENDENT VARIABLE—is the number of times a child hits, kicks, or pushes another child during a 15-minute play session. Using a population of children of a certain age (say, 6 to 7 years old), children are randomly

Table 1 Case of Significant Main Effects and No Significant Interaction

Source	DF	SS	MS	F	p
Gender	1	122.51	122.51	33.56	<.0001
Game	1	103.51	103.51	28.35	<.0001
Gender $\times$ Game	1	1.01	1.01	0.28	.600

assigned to a condition where they play a 30-minute violent video game or a 30-minute nonviolent video game before the play session with other children. This constitutes the independent variable manipulation. This RANDOM ASSIGNMENT is done separately for boys and girls, resulting in a 2×2 design with two levels of type of video game crossed with two levels of gender.

The resulting data and two-way ANOVA then can be used to address the following research questions: (a) Are boys more aggressive than girls? (b) Are children more aggressive following experience with a violent video game? and (c) Do boys and girls differ in the effect of exposure to violent video games? In ANOVA parlance, the first two questions are addressed by MAIN EFFECT tests—that is, averaged over experimental conditions, are the mean aggression scores significantly higher for boys than for girls, and, averaged over gender, are the mean aggression scores significantly higher in the Violent Video Game condition than in the Nonviolent Video Game condition? The third question, which is probably the one of greatest interest, is addressed by a test of INTERACTION—does the effect of exposure to violent

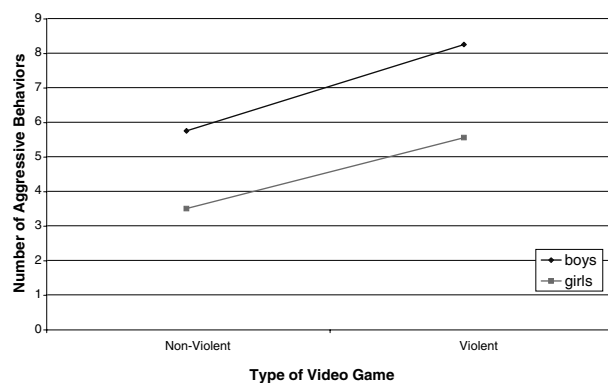
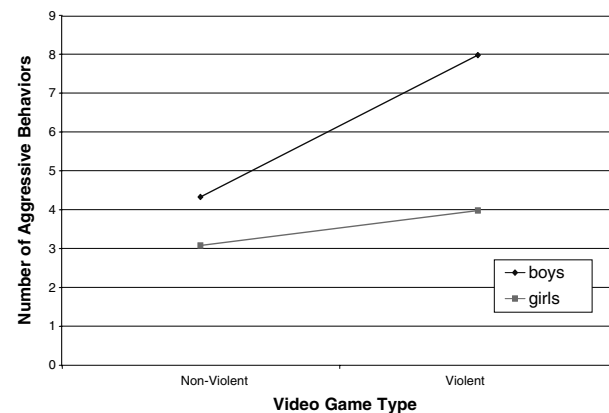
Table 2 Case of Significant Main Effects and Significant Interaction

Source	DF	SS	MS	F	p
Gender	1	137.81	137.81	50.76	<.0001
Game	1	103.51	103.51	38.12	<.0001
Gender $\times$ Game	1	37.81	37.81	13.93	<.0004

video games differ significantly between boys and girls?

DISPLAYING THE RESULTS

These three tests—the two main effects and the interaction—are independent of each other in a two-way ANOVA, meaning that any combination of significant and nonsignificant results is possible (see Hays, 1994; Jaccard, 1998; and Levin, 1999, for computational procedures). An ANOVA table, coupled with a graph of the mean results for each combination of Factors A and B, gives the complete picture. A couple of interesting sets of hypothetical results are illustrated above. In the first example, the ANOVA table shows that both main effects are significant, whereas the interaction is not (see Table 1). The graph, showing approximately parallel lines for boys and girls, illustrates that boys are more aggressive than girls, but that children of both sexes show comparable increases in aggression following exposure to a violent video game compared to a nonviolent video game (see Figure 1). In the second example, both main effects are again significant, but so is the interaction (see Table 2).

**Figure 1** Aggressive Behavior as a Function of Gender and Video Game Type**Figure 2** Aggressive Behavior as a Function of Gender and Video Game Type

The accompanying graph shows that boys are much more strongly influenced by exposure to violent video games than are girls, as illustrated by the greater separation of lines for the Violent condition than for the Nonviolent condition (see Figure 2). Detecting and describing the interaction is what enables us to distinguish between these two sets of results. The use of two-way ANOVA in conjunction with a factorial study allows the researcher to investigate the separate and combined effects of two variables in a single study. The same principle can be extended to three-way ANOVA with three variables, and so forth.

—Irwin P. Levin

REFERENCES

- Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.
- Jaccard, J. (1998). *Interaction effects in factorial analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–118). Thousand Oaks, CA: Sage.
- Levin, I. P. (1999). *Relating statistics and experimental design: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–125). Thousand Oaks, CA: Sage.

TYPE I ERROR

Type I and Type II errors are errors that can occur with HYPOTHESIS TESTING. Even though data are available only for a RANDOM SAMPLING of observations and not on the entire POPULATION, the sample data can be used to draw conclusions about the larger population. This means extrapolation from the small sample to the large population, and therefore, we may make errors. The two main types of errors are called errors of Type I and Type II.

The extrapolation from sample to population is done by setting up a NULL HYPOTHESIS. There, we specify one or more values for a population parameter. These are examples of null hypotheses: $H_0 : \mu = 100$ (the population MEAN equals 100), $H_0 : \sigma^2 \geq 25$ (the population VARIANCE is equal to or greater than 25), $H_0 : \beta = 0$ (the population REGRESSION COEFFICIENT equals 0). We want to learn about the values of the population parameters, but because the values are not known, a null hypothesis becomes more of a question. For the first null hypothesis, could it be that the population mean equals 100?

Consider the case where the null hypothesis really is true. If the sample data are such that we reject the null hypothesis and conclude that it is not true, then we have made a mistake. Because the null hypothesis really is true, it should not have been rejected. This is a Type I error—rejecting a null hypothesis even though it really is true.

We will never know whether we commit this error or not, because the true value of the population parameter remains forever unknown. All we do here is examine what happens when the null hypothesis is true and we have data that lead us to reject the null hypothesis. It is possible to speculate on how often we will make such a Type I error. Suppose, over a lifetime, we examine 1,000 different null hypotheses, where 800 of them are false and 200 are true. If we commit the error of Type I 5% of the time, we will have made this error in 10 of the 200 true cases (10 of 200 is 5%). The 800 other null hypotheses are irrelevant here, because we deal only with those null hypotheses that are true.

To decide whether we reject a null hypothesis or not, we use the sample data to compute the so-called P VALUE. Most statistical software will compute this number for us. It tells us, in the case where the null hypothesis is true, how often we get our sample data or more extreme data. For example, let $H_0 : \beta = 0$, and from the data, $b = 4.56$ with $p = .001$. Thus, if the population regression line is truly horizontal (there is no relationship between the two variables), we will get a sample regression coefficient of 4.56 or larger only one time in 1,000 replications of this study.

—Gudmund R. Iversen

REFERENCES

- Iversen, G., & Gergen, M. (1997). *Statistics: The conceptual approach*. New York: Springer-Verlag.
- Moore, D. S., & McCabe, G. P. (1998). *Introduction to the practice of statistics*. New York: W. H. Freeman.
- Utts, J. M. (1996). *Seeing through statistics*. Belmont, CA: Duxbury.

TYPE II ERROR

STATISTICAL INFERENCE methods are used to draw conclusions about values of unknown PARAMETERS. In HYPOTHESIS TESTING, a NULL HYPOTHESIS is set up specifying one or more values of the parameters: $H_0 : \mu = 100$ (the population MEAN equals 100),

$H_0 : \sigma^2 \geq 25$ (the population VARIANCE is equal to or greater than 25), or $H_0 : \beta = 0$ (the population REGRESSION COEFFICIENT equals 0). Because the parameter values are unknown, a null hypothesis becomes a statement about possible values of the parameters. It is as if we are asking whether the parameters could have the values specified in the null hypotheses.

SAMPLE data are used to answer questions posed in the null hypotheses. Based on the data, we make decisions about whether to reject or not reject the null hypothesis. Because such a decision is an extrapolation from the limited sample data to a larger population, it is possible to make errors. If the actual value of the parameter is the value specified in the null hypothesis, then the null hypothesis is true. If we still reject the null hypothesis, we make a TYPE I ERROR.

It may be, however, that the actual value of the parameter is not equal to the value specified in the null hypothesis. In that case, the null hypothesis is false and should be rejected. But if the data are such that we fail to catch the fact that the null hypothesis is false, then we have made an error. This is a Type II error—not rejecting a false null hypothesis.

Whether such an error is committed depends on two things. First, it depends on how far the true value of the parameter is from the value specified in the null hypothesis. If the true population mean $\mu = 101$, and the null hypothesis states that the population mean $\mu = 100$, then the null hypothesis is false and should

be rejected. But with such a small difference between the two, it will be very difficult to discover that the null hypothesis is false. In many replications of the study, the sample means will cluster around 101 rather than 100. However, this difference may be so small that it does not matter for the problem at hand.

Second, whether the data will detect that the null hypothesis is false will depend on how many observations there are in the sample. With small samples, it is more difficult to discover that a null hypothesis is false than it is with large samples. With large samples, the sample statistic in replications of the study will cluster more tightly around the true parameter value, and we reject the null hypothesis more often. For a fixed value of the parameter and a given sample size, we can compute how often (the PROBABILITY) we will get data such that the null hypothesis is erroneously rejected. The *power* of the test is defined as one minus this probability.

—Gudmund R. Iversen

REFERENCES

- Iversen, G., & Gergen, M. (1997). *Statistics: The conceptual approach*. New York: Springer-Verlag.
- Moore, D. S., & McCabe, G. P. (1998). *Introduction to the practice of statistics*. New York: W. H. Freeman.
- Utts, J. M. (1996). *Seeing through statistics*. Belmont, CA: Duxbury.

U

UNBALANCED DESIGNS

Unbalanced designs (sometimes called nonorthogonal designs) are FACTORIAL research designs in which QUALITATIVE INDEPENDENT VARIABLES are correlated. This circumstance, akin to MULTICOLLINEARITY in REGRESSION analysis, makes impossible the unambiguous estimation of the correlated independent variables' effects. As a rule, designs are unbalanced when sample sizes are unequal—that is, when the numbers of subjects per cell differ. A special case occurs when one cell is entirely empty.

Yates (1934) showed that in analyzing data from an unbalanced design, ordinary ANALYSIS OF VARIANCE (ANOVA) will, for most purposes, be misleading. Using Type I sums of squares (SS), ANOVA assigns explanatory weight to an independent variable partly on the basis of that variable's sequence in the model. The first variable captures all of the variance confounded with it, whether that variance is unique or is shared with one or more other independent variables. The independent variable that follows the first will be given weight only to the extent that it adds an additional increment to the explained SS, and so on with subsequent variables.

To address this arbitrary feature of Type I SS, Yates developed a method called weighted squares of means, now generally known as Type III SS, which has become the routine analytic choice with unbalanced data. In that method, SS are assigned to each independent variable as if it were the last variable in a Type I model; that is, a variable's weight is determined by its unique contribution to the explanation after all other factors

have been taken into account. Thus, an independent variable's SS is in fact its partial SS. Note that the SS associated with the confounded independent variables is left unassigned here.

Nelder and Lane (1995) find this approach unrealistic, arguing that it may, among other things, produce models with INTERACTIONS but lacking the corresponding MAIN EFFECTS. In its place, they propose the use of Type II analysis (also known as Yates's method of fitting constants). The Type II method involves adjustment for the other factors only if those factors do not include the variable being estimated. Thus, X_1 would not be adjusted for $X_1 \cdot X_2$ because the latter contains the former. When a model includes only main effects, Type II analysis is no different from Type III. In other cases, Type II requires choices about which factors to drop from the model to proceed.

When one or more cells in a factorial model are empty, the results from Type I, II, or III analysis often defy interpretation. A fourth method, more technically complex, is available. Type IV SS involves constructing a hypothetical, balanced data cell matrix that then is used as the basis for estimating effects. The solution may not be unique, and for that reason and others, great caution is urged in its interpretation.

—Douglas Madsen

See also MODEL I ANOVA

REFERENCES

- Nelder, J. A., & Lane, P. W. (1995). The computer analysis of factorial experiments: In memoriam—Frank Yates. *The American Statistician*, 49, 382–385.

Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association*, 29, 51–66.

UNBIASED

An ESTIMATOR is said to be unbiased when the expected value of the estimate $\hat{\beta}$ equals the true value of the PARAMETER β : $E(\hat{\beta}) = \beta$. Unbiasedness is a desirable property for an estimator because it provides confidence that our estimates of a parameter, on average, equal the true parameter. An estimate is biased when the expected value of the parameter does not equal the true value $E(\hat{\beta}) \neq \beta$. The BIAS can be calculated by subtracting the true value of the parameter from the expected value of the estimate: $\text{bias} = E(\hat{\beta}) - \beta$.

One way to think of unbiased estimators is by imagining two people shooting a bow and arrow and trying to hit the center of a target. Each person is a different estimator, and the center of the target represents the true value of the parameter being estimated. Person 1 shoots 100 arrows; the average arrow shot hits the center of the target. Person 2 shoots 100 arrows. However, Person 2's aim is not as good, and the arrows go slightly to the left. The average arrow shot by Person 2 is about 1 foot to the left of the center. The average locations of the arrows represent the expected values of the estimates generated by the two estimators. Person 1 would be an unbiased estimator because, on average, he or she was able to hit the center. Person 2 would be a biased estimator because the average arrow went to the left of the true value.

It should be noted that unbiasedness is a property of the PROBABILITY DISTRIBUTION generating the estimates. An unbiased estimating technique will not give us the true value of our parameter on each try. Rather, it is only after repeated SAMPLING that the average value of the estimated parameter is equal to the true value of the parameter.

Just because an estimator is unbiased does not mean that it is the ideal estimator (Wooldridge, 2002). Rather, an unbiased estimator does us little good if it is not an efficient estimator (see EFFICIENCY). An inefficient estimator's probability distribution has a large variance, causing high uncertainty in the probability of any given estimate equaling the true value of the parameter.

Occasionally, we will use estimators that are biased because they have other desirable properties.

Unbiasedness is an attractive feature of ORDINARY LEAST SQUARES (OLS) as an estimation technique. Wooldridge (2002) explains that OLS is unbiased when four conditions are met. First, the relationship between the dependent and independent variables is LINEAR in functional form. Second, the observations are drawn from a RANDOM SAMPLE. Third is the ASSUMPTION of zero conditional mean. Zero conditional mean states that, given x , the expected value of the errors is zero. Fourth, there is variation in the variables being studied. When these conditions are met, OLS is an unbiased estimator. When an estimator is shown to be efficient, has a minimum variance, and is unbiased, it is a special class of estimator called a BEST LINEAR UNBIASED ESTIMATOR (BLUE).

—Heather L. Ondercin

REFERENCE

Wooldridge, J. M. (2002). *Introductory econometrics: A modern approach*. Cincinnati, OH: South-Western College Publishing.

UNCERTAINTY

Social scientists are doubly concerned with uncertainty: Formal research methods are designed to quantify the uncertainty that arises when building THEORIES of an imperfectly predictable social world. But much of the variability in human behavior that demands these methods is generated by the judgments that individual society members make in light of their own uncertainty.

Rational expectations theories in economics are a rare case when the two concerns are explicitly related: The price of a stock, a society-level phenomenon, is predicted to evolve randomly over time because all the available systematic information will have been used by individual investors to minimize their uncertainty about its value. Any remaining price movement reflects residual investor uncertainty.

PROBABILITY provides the basis for addressing both concerns with uncertainty: Most economic and political theory assumes that probabilistic inference is a reasonable BEHAVIORAL model of individual decision making in an uncertain world. Uncertainty about

particular social scientific theories is also expressed using probabilities and the associated machinery of STATISTICS. But probability is not the only possible choice for quantifying uncertainty, so it is worth considering why it is a good one.

There are several arguments motivating individuals to use probability theory for managing their uncertainty. “Dutch Book” arguments show that individuals who act on their uncertainty in a way that is inconsistent with probability theory can, in gambling situations, always be fleeced (Ramsey, 1960). A more general motivation for probability was provided by Richard Cox (1946/1990). Cox showed that for any numerical measure of certainty, plausibility, or confidence in a Proposition A that we might invent, it must be a simple rescaling of the probability that A is true if it has the following intuitive properties:

1. Transitivity: If A is more plausible than B, and B is more plausible than C, then A is more plausible than C.
2. The plausibility of A is a function of the plausibility of not-A. (For example, the more plausible A is, the less plausible not-A is.)
3. The plausibility of A and B together is a function of the plausibility of A, as well as the plausibility of B when A is certain.

If these properties apply to the measure, then the number we assign to A will be a constant multiple of the probability of A. If Principles 1 through 3 are accepted as minimum requirements, then probability theory is the correct way to quantify uncertainty. Since Cox’s argument does not depend on whether the uncertainty being quantified is personal or social scientific, it provides a general motivation for quantifying uncertainty with probability. Despite the wide application of rational expectations theories in economics, most psychologists doubt that probability theory is a good descriptive model of individual reasoning under uncertainty (Tversky & Kahneman, 1974). For example, people are overconfident relative to available data and partially neglect the prior probability of events when they make predictions (base-rate neglect). These psychological facts are not necessarily problematic for rational expectations theories because individual expectations need only be consistent with probability theory on average. Nor do they prevent probability-based theories of personal inference; people may be making a probabilistically correct inference on an

incorrect internal model or taking into account causal constraints (Cheng, 1997). People also seem to process information in a way that trades some of the advantages of probabilistically correct INFERENCE for speed by using an incorrect but less computationally demanding inference process that is only approximately correct (Gigerenzer & Goldstein, 1996).

The question of whether scientific uncertainty should be quantified in the same way as personal uncertainty defines the debate in statistics between frequentists and BAYESIANS. We focus on this question with two interpretations of a familiar example.

A LINEAR REGRESSION model of the relationship between some quantity Y and two potentially explanatory variables X_1 and X_2 can be written as

$$Y = X_1b_1 + X_2b_2 + e$$

$$e \sim \text{Normal}(0, \sigma^2). \quad (1)$$

In a Bayesian treatment of this model, parameter values are uncertain, and this uncertainty is described by a PRIOR PROBABILITY distribution, $p(b_1, b_2, \sigma^2)$. But even if the parameters were known with certainty, the outcome Y would be uncertain. This uncertainty is factored into a DETERMINISTIC part, the additive relation $X_1b_1 + X_2b_2$, and a STOCHASTIC part by assuming a particular probability distribution for the RANDOM VARIABLE e . Uncertainty about Y can be partly reduced by observing values of X_1 and X_2 , down to the minimum level determined by e ; e may represent the work of a truly random social process or our belief that there are variables relevant to Y not included in the model. Equation (1) defines a distribution over outcomes $p(Y|b_1, b_2, \sigma^2)$ that is conditional on particular values of the PARAMETERS and explanatory variables (ignoring the known explanatory variables). BAYES THEOREM combines these two distributions into a posterior probability distribution, $p(b_1, b_2, \sigma^2|Y)$. The posterior probability distribution combines uncertainty about parameter values and uncertainty about the value of Y that would persist even if everything else were known. The peak of the posterior is the most probable value for the parameters, and a 95% interval has a 0.95 probability of containing the true value.

The posterior distribution is conditional on Y because after the data are observed, their values are known with certainty. Conditioning on the observed data is only possible because there is a prior distribution. For frequentists, it is inappropriate to use prior distributions in science because it reduces the objectivity

of scientific inference. In this case, probability is restricted to describing objectively random mechanisms, and uncertainty is confined to states of complete ignorance. Under this interpretation, e describes an objective physically or socially random process, and the parameters are fixed but completely unknown.

To infer values for the parameters, frequentists choose a function of the available Y values to be an ESTIMATOR and then compute a POINT ESTIMATE and CONFIDENCE INTERVAL. Because Y has a probability distribution in virtue of containing e , parameter estimates (but not the parameters themselves) also have a distribution because they are functions of the objectively random Y . Confidence intervals reflect uncertainty by using probability only indirectly: It is the method of interval construction that is associated with a probability of, say, 0.95 in the sense that in repeated sampling, 95% of intervals constructed this way would contain the true parameter value. Because observed Y values cannot be conditioned on, a confidence interval depends on infinite numbers of hypothetical outcomes—values of Y that did not actually occur but could have done.

Although a definition of probability in terms of repeated samples allows frequentists to avoid quantifying levels of uncertainty, it can be problematic. For example, much social science research is concerned with explaining essentially unique situations in which the idea of a repeat sample may not make sense. In contrast, interpreting e as an uncertainty measure is consistent with any position on the possibility of repeated sampling because there will be uncertainty about outcomes in either case.

As for the uncertainty represented by priors over parameters, it is important to see that a posterior distribution states what it would be rational to believe on the basis of observed data given a set of prior beliefs. Given a different set, the rational beliefs might well be different. Although the prior is inevitably subjective, it is also explicitly represented and introduces no ambiguity into inference, and the consequences of changing it may be informative about the information content of the data.

The debate between frequentists and Bayesians about social scientific inference turns on whether probability is always the best representation of uncertainty—in other words, whether the benefits of being able to condition on data that are actually observed, removing sampling theoretic constraints

and being consistent with foundational arguments for probability as a measure of uncertainty should outweigh concerns about introducing subjective elements into the process of increasing scientific knowledge about the social world.

—Will Lowe

REFERENCES

- Berger, J. O., & Wolpert, R. L. (1984). *The likelihood principle: A review, generalizations, and statistical implications*. Hayward, CA: Institute of Mathematical Statistics.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367–405.
- Cox, R. T. (1990). Probability, frequency, and reasonable expectation. In G. Shafer & J. Pearl (Eds.), *Readings in uncertain reasoning* (pp. 353–365). New York: Morgan Kaufmann. (Original work published 1946)
- Edwards, A. W. F. (1972). *Likelihood: An account of the statistical concept of likelihood and its application to scientific inference*. Cambridge, UK: Cambridge University Press.
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650–669.
- Ramsey, F. P. (1960). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays*. New York: Harcourt Brace.
- Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.

UNIFORM ASSOCIATION. See
ASSOCIATION MODEL

UNIMODAL DISTRIBUTION

This is a DISTRIBUTION of values for a single VARIABLE that only has one mode and a single “peak.” The mode is one of three common measures that are used to determine a variable’s “typical score” and is the most frequently occurring value of any given variable (Lewis-Beck, 1995, p. 8).

An example of a unimodal distribution is the standard NORMAL DISTRIBUTION. This distribution has a MEAN of zero and a STANDARD DEVIATION of 1. In this particular case, the mean is equal to the MEDIAN and mode. Moreover, the standard normal distribution

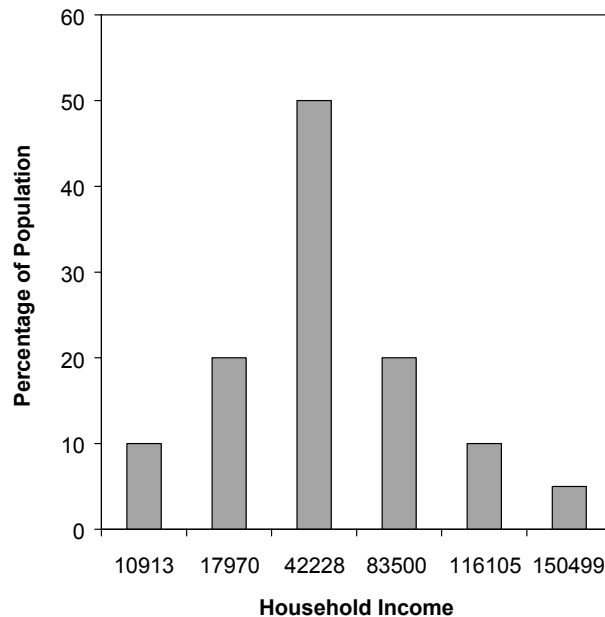


Figure 1 Histogram for Household Income at Selected Percentiles

only has a single, equal mean, median, and mode. Therefore, it is a unimodal distribution because it only has one mode.

However, a unimodal distribution does not have to have SYMMETRY like the standard normal distribution. A unimodal distribution can also be SKEWED to the left or right. An example of such a distribution is that of income in the United States.

Figure 1 illustrates the annual income of the U.S. population by different categories (DeNavas-Walt & Cleveland, 2002, p. 20). The MEDIAN income, in which half of the values are greater and half are smaller, falls in the \$42,228 category. This figure suggests that the distribution of American household income is somewhat skewed to the left because the median is closer to the bottom 10% (at \$10,913) than to the top 10% of all households (at \$150,499). Furthermore, the modal category is \$42,228. There is a much larger percentage of households (about 50%) in this income category than in the other income categories, indicating the distribution is clearly unimodal.

Regardless of whether a unimodal distribution is skewed or symmetric, it must have a single “peak.” A distribution with a single “peak” is one in which there is a single most frequently occurring value of any given variable, as illustrated by its FREQUENCY DISTRIBUTION. If a distribution has two “peaks,” then it is a BIMODAL DISTRIBUTION. Finally, if a distribution

has more than two “peaks,” then it is a MULTIMODAL DISTRIBUTION.

—Kenneth W. Moffett

REFERENCES

- DeNavas-Walt, C., & Cleveland, R. W. (2002). *Money income in the United States* (United States Census Bureau, Current Population Reports, P60–218). Washington, DC: Government Printing Office. Available: www.census.gov/prod/2002pubs/p60-218.pdf
- Kennedy, P. (1998). *A guide to econometrics* (4th ed.). Cambridge: MIT Press.
- Lewis-Beck, M. S. (1995). *Data analysis: An introduction*. Thousand Oaks, CA: Sage.

UNIT OF ANALYSIS

A unit of analysis is the most basic element of a scientific research project. That is, it is the subject (the who or what) of study about which an analyst may generalize. In the social sciences, countries, international alliances, schools, communities, interest groups, and voters are often the units of analysis in studies of economic development, war, teaching strategies, social capital, policy outcomes, and vote choice.

Units of analysis may be different from the units of observation. For example, political scientists may aggregate survey data at the country level by estimating the mean score on an item measuring political participation, assigning that same score to the country to identify how patterns of participation differ cross-nationally. In this case, the unit of observation is the individual survey respondent, but the unit of analysis is the country. Likewise, achievements of students are often aggregated to allow schools to be compared. It is the country and the school about which the analyst generalizes, not the citizens or students (see AGGREGATION for a word of caution).

In these examples, more detailed information is available than the analyst requires for his or her project. Often in the social sciences, however, this is not the case: Researchers frequently examine the behavior of units that are unobserved. For example, following Achen and Shively (1995), an analyst may wish to examine the relationship between a social attribute and voting behavior but has only data aggregated at the district level available. That is, he or she has the proportions of those living in the district who have this attribute and of those supporting the party. If both

proportions are large, the analyst is tempted to infer that there is an association between these variables at the level of the individual: Voters with this attribute are likely to vote for the party. This type of cross-level inference, in which the unit of analysis does not coincide with the unit of observation, is misleading and should be approached with caution (see ECOLOGICAL FALLACY).

Even when units of analysis match the units of observation, however, the characteristics of social units complicate scientific inference. Boucke (1923) notes two differences that distinguish social units from the units of analysis common in the natural sciences: First, the units of social science research are not measurable in a straightforward way or on a determinate scale. Often, a considerable proportion of variance in social phenomena can be attributed to measurement error (see SYSTEMATIC ERROR). Also, some units of interest may be observable only in their causal implications (see Blalock, 1961). Second, as evident in the examples presented above, units of analysis in the social sciences can usually be divided into various subunits: Communities may be divided into families or households, geographic neighborhoods, or individual members. Boucke concludes that units of analysis in social science are not “things” but the relationships or networks connecting the families, neighbors, or, at the most basic level, the interests of individuals. Subunits rarely have the potential for such substantive meaning or CONFOUNDING effects in the natural sciences. Furthermore, the variability of relations between subunits makes scientific inference a difficult task: The composite nature of the units of analysis in social research makes “predictions risky, and the more so, the more fully they concern psychic rather than material relations” (Boucke, 1923, p. 460). Analysts working in the social sciences would do well to accommodate these implications of their choice of unit of analysis in their research designs.

—Karen J. Long

REFERENCES

- Achen, C., & Shively, W. P. (1995). *Cross-level inference*. Chicago: University of Chicago Press.
- Blalock, H. M. (1961). Theory, measurement and replication in the social sciences. *American Journal of Sociology*, 66(1), 342–347.
- Boucke, O. F. (1923). The limits of social science: II. *American Journal of Sociology*, 28(4), 443–460.

UNIT ROOT

A times series can be modeled as a function of its own past, that is, as an autoregression. In its simplest form, such a series is written as

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t,$$

where the ε_t s are assumed to be independent of each other. This series has a unit root if $\beta = 1$. The presence of a unit root has critical consequences for interpretation and ESTIMATION, both for the simple autoregression and for a more complex model with a lagged dependent variable and other independent variables. It is critical to test for whether an autoregressive series has a unit root and, if so, to take account of this in estimating the model parameters.

If a series has a unit root, then we can rewrite the autoregressive form as $y_t = \alpha + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \dots$, so that current y is the sum of all previous errors. (Integration is just a form of summation, which is why such a series is called “integrated.”) Because the errors are independent, the variance of y grows larger over time. Thus, all the usual assumptions that make the asymptotic theory of REGRESSION work no longer hold and all standard methods are incorrect. In particular, if the autoregressive series has a unit root, but we try to estimate β by ORDINARY LEAST SQUARES, we will underestimate β and misestimate its standard error. Interpretation is also dramatically affected: With a unit root, small perturbations of a series will persist forever rather than die out, as would happen with a stationary series (where $|\beta| < 1$). The CENTRAL LIMIT THEOREM does not even hold for integrated series, and so all standard distribution theory breaks down in this case.

Dickey and Fuller were the first to design a correct test for the null hypothesis that a time series has a unit root. This consists of using ordinary least squares to estimate β and then forming the Dickey-Fuller statistic $(1 - b)/SE$, where b is the ordinary least squares estimate of β and SE is the ordinary least squares standard error of this estimate. With a stationary time series, this statistic has a τ -DISTRIBUTION; under the null hypothesis of a unit root, Dickey and Fuller had to use deep mathematics to derive the appropriate distribution of the statistic. Thus, users compute the Dickey-Fuller statistic and compare its value to the

distribution tabulated by Dickey and Fuller. There are many modern improvements on this test, but they are all based on the original idea of Dickey and Fuller. When using the Dickey-Fuller test, it is important to remember that the null hypothesis is that the series is integrated. If this null hypothesis is rejected, then standard methods can be used. If it is not rejected, this does not mean that the series has a unit root—it only means that we do not have enough information to reject this hypothesis. Unfortunately, there is no test that reverses this logic, so researchers typically assume a unit root if the Dickey-Fuller test fails to reject this hypothesis.

If a series has a unit root, then standard ordinary least squares methods cannot be used. One solution is to take first differences and model those because the first differences will usually not show a unit root. Substantively, this is equivalent to modeling only the short-run behavior of a series, ignoring any long-run behavior of that series. Alternatively, analysts faced with a unit root can use the methodology of COINTEGRATION.

—Nathaniel Beck

REFERENCES

- Engle, R. F., Granger, C. W. J. (1991). *Long run relationships: Readings in cointegration*. New York: Oxford University Press.
- Mills, T. C. (1990). *Time series techniques for economists*. Cambridge, UK: Cambridge University Press.

UNIVARIATE ANALYSIS

Univariate analysis is the analysis of the statistical characteristics of a single VARIABLE, including its DISTRIBUTION, its CENTRAL TENDENCY, and its SPREAD.

The first step in univariate analysis is to plot the observations in a table or graph to examine the variable's distribution. NOMINAL variables can be presented in a PIE CHART or BAR chart. In a pie chart, each wedge in a pie represents a category of the variable. In a bar chart, the height of each bar represents the number or proportion of observations in each corresponding category of the variable. INTERVAL and RATIO data are often presented in a HISTOGRAM, BOXPLOT, or QUANTILE plot. Some variables have SYMMETRIC distributions, whereas other variables have SKEWED

distributions with OUTLIERS that can distort STATISTICS that are calculated on them.

Summary statistics for a single variable are often called DESCRIPTIVE STATISTICS. Central tendency measures summarize the typical value of the variable. The arithmetic average, known as the MEAN, is usually used as a measure of central tendency for numeric variables. The MEDIAN—the value of the variable for the middle ordered observation—is used for ORDINAL variables. The MODE—the value that occurs most often—is used for nominal data. The median is often used in place of the mean for numeric data when there are outliers in the data or when the variable's distribution is extremely skewed.

Measures of DISPERSION indicate the amount of variation in a variable's observations. For nominal or ordinal data, crude measures of dispersion include the number of categories and the number of observations in the modal category. For numerical variables, the simplest measure of dispersion is the RANGE, the difference between the highest and lowest observation. Because outliers can distort the range, an alternative is the INTERQUARTILE RANGE, the difference between the 25th percentile and the 75th percentile. The interquartile range is prominently featured in boxplots. For interval and ratio data, the STANDARD DEVIATION and VARIANCE (the square of the standard deviation) are common measures of dispersion. Larger values of the standard deviation and variance indicate more dispersion among the observations.

The distribution of a single variable can also be examined to determine how closely it approximates a symmetrical, bell-shaped NORMAL DISTRIBUTION. Symmetrical distributions have the same shape on either side of the median. Skewness statistics indicate how much a distribution deviates from being symmetrical. The KURTOSIS statistic indicates the extent to which a distribution has a tall, thin peak or a short, flat peak around the mode.

—David C. Kimball and Herbert F. Weisberg

REFERENCES

- Fox, W. (1998). *Social statistics* (3rd ed.). Bellevue, WA: MicroCase.
- Jacoby, W. G. (1997). *Statistical graphics for univariate and bivariate data*. Thousand Oaks, CA: Sage.
- Lewis-Beck, M. S. (1995). *Data analysis: An introduction*. Thousand Oaks, CA: Sage.

Weisberg, H. F. (1992). *Central tendency and variation*. Newbury Park, CA: Sage.

UNIVERSE. See POPULATION

UNOBSERVED HETEROGENEITY

Unobserved heterogeneity refers to latent or hidden characteristics of observations. These characteristics are properties of individuals or subgroups that cannot be measured or are left unmeasured in a study. Ordinarily, individual or group characteristics are incorporated into statistical models as COVARIATES (or INDEPENDENT VARIABLES). When unmeasured characteristics affect the process under investigation, unobserved heterogeneity among observations is said to exist.

For some statistical models, failure to account for heterogeneity among individuals or groups can lead to BIASED PARAMETER ESTIMATES, misinterpretation of results, or SPURIOUS time-covariate interactions. Consequently, statistical techniques have been developed that explicitly incorporate unmeasured heterogeneity as the component of a model. The result is one or more additional parameters that describe how the unmeasured characteristics are distributed in the population from which the sample was drawn.

AN EXAMPLE

Consider the ages at marriage in a rural area of Bangladesh (Figure 1). The heavy line shows the distribution that arises when gender is not considered (or left unmeasured). Suppose a lognormal distribution is believed to be a proper model for age at marriage. Fitting the observations to a lognormal distribution (without reference to sex) yields parameters that give a MEAN age (STANDARD DEVIATION) of 23.3 (5.9) years. Failure to consider gender masks substantial subgroup differences. In fact, a gender-specific analysis yields a mean age of marriage of 27.3 (5.5) years for males and 19.5 (3.4) years for females. The subgroup differences are so large that parameter estimates for any other measured covariate (say, the effect of education on age at marriage) are likely to be BIASED by the lack of information on gender.

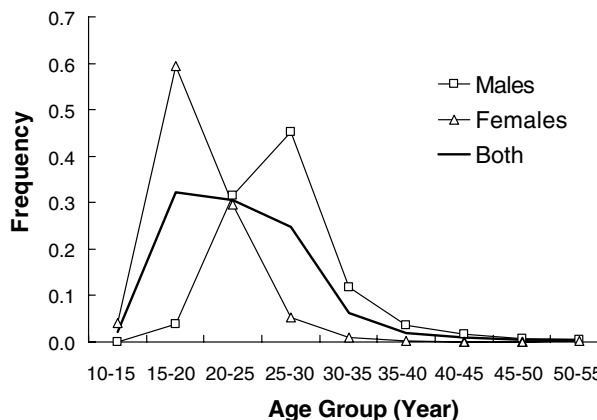


Figure 1 Ages at Marriage in Matlab, Bangladesh, in 1985

SOURCE: Data are from the International Centre for Diarrhoeal Disease Research, Bangladesh (1992).

With no information on gender, one solution is to estimate parameters for a finite mixture model. A four-parameter mixture model can be estimated assuming equal numbers of males and females and yields estimates of 25.4 (7.1) years for one subgroup and 21.4 (4.0) years for a second subgroup. The subgroups correspond to males and females, respectively.

“DEVIOUS DYNAMICS” IN HAZARDS MODELS

Failure to account for unobserved heterogeneity can lead to serious problems of interpretation for survival analysis and related methods that deal with durations to events (hazards models or EVENT HISTORY models). For these models, unobserved heterogeneity refers to unmeasured differences among the units of observation in their risk of “failure” (i.e., risk of experiencing the event under investigation). Ignoring unmeasured heterogeneity in risk of failure leads to a downward bias in risk with time.

Consider mortality risk in infants, which is initially very high and declines rapidly over the first years of life. Taken literally, this observation implies that infants undergo reverse senescence: Individuals “get better” with age. In part, this pattern reflects known biological processes such as maturation of the immune system. On the other hand, the same “infant mortality” pattern is seen for other complex systems such as computers and automobiles—it is less plausible to believe

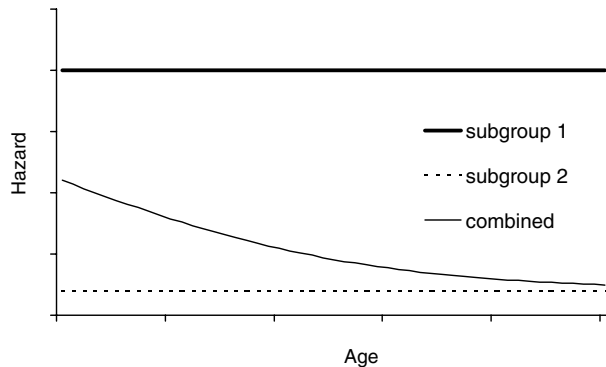


Figure 2 A Mixture of Two Homogeneous Subgroups, Each at a Constant Hazard (or Risk) of Failure

NOTE: Risk appears to decline with age when the subgroups are combined.

that individual computers experience a decline in risk of failure during early service. The apparent reverse senescence is an artifact of the changing composition of the sample over time, a product of the “devi-ous dynamics of aging cohorts” (Vaupel & Yashin, 1985a, 1985b).

Figure 2 shows how aggregate risk is biased downward with time as a result of unobserved heterogeneity. Two subgroups are shown, one at a high and constant risk of failure and a second at low and constant risk of failure. The subgroups initially contain the same number of individuals. When the subgroups are combined and treated as a single homogeneous population, a decline in risk of failure is observed with age. This decline occurs because individuals in the high-risk subgroup are failing at a higher rate. Over time, the surviving population is increasingly made up of low-risk individuals.

An important special case of heterogeneity arises when all individuals in one of the latent subgroups are not at risk of failure. For example, a demographer studying fecundability in couples might observe times from marriage until each couple’s first birth. Some fraction of the couples may be sterile and will therefore never give birth. Difficulties arise when some of the fecund couples drop out of the study or are lost to follow-up before their first child is born (i.e., they are right censored; see CENSORING). Those couples whose first birth is not observed come from two fundamentally different subgroups: the right-censored fecund individuals and the sterile individuals. Special

analytical models—called sterility models, immune models, or mover-stayer models—can be used to estimate an unbiased distribution of waiting times to first birth and the initial fraction of sterile individuals. This type of process arises in many social sciences contexts, including studies of marriage, divorce, birth spacing, recidivism, occupational mobility, and residential mobility.

FRAILTY MODELS

Unobserved heterogeneity in risk of failure can be distributed among individuals in a way that approximates a continuous distribution. An important illustration is the concept of *frailty* that developed out of demographic studies of mortality (Vaupel, Manton, & Stallard, 1979). Individuals have unmeasurable (or unmeasured) genetic, environmental, or behavioral liabilities or *frailties* that uniquely modify their risk of mortality from some baseline distribution. An individual frailty is described by the quantity z_i , but because z_i is not measurable, an alternative is to estimate parameters for a probability distribution that describes how z is distributed among individuals.

For many types of frailty models, the effect of z on the hazard of failure is specified as a proportional hazards model, so $h(t|z) = h(t)z$ for time t . The frailty distribution is selected to have a mean of 1; a variance parameter is estimated and describes the way frailty is distributed in the population. The most commonly used parametric frailty distribution is the gamma distribution. Occasionally, the specification $h(t|z) = h(t)e^z$ is used when the distribution of z has a mean of zero; the NORMAL DISTRIBUTION is frequently employed.

Results from a frailty model can be sensitive to the choice of frailty distribution. An alternative approach specifies the distribution of heterogeneity as a piecewise discrete density function (Heckman & Singer, 1982). The number and location of mass points are determined empirically. This approach may be particularly appropriate when theory does not suggest an appropriate distribution for z . Even so, this strategy can be sensitive to the parametric model used for the hazard function (see Trussell & Rodríguez, 1990).

In addition to frailty models, there are many other commonly used parametric models of unobserved heterogeneity. For example, the BINOMIAL and GEOMETRIC distributions can incorporate a distribution of heterogeneity for the probability (p) of the event under investigation. A beta distribution is frequently used as

the distribution of p , if only because the range of the distribution lies from 0 to 1.

—Darryl J. Holman

See also HETEROGENEITY

REFERENCES

- Heckman, J., & Singer, B. (1982). Population heterogeneity in demographic models. In K. Land & A. Rogers (Eds.), *Multi-dimensional mathematical demography* (pp. 567–599). New York: Academic Press.
- Heckman, J., & Singer, B. (1994). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 62(2), 271–320.
- International Centre for Diarrhoeal Disease Research, Bangladesh (ICDDRDB). (1992). *Demographic surveillance system—Matlab: Registration of demographic events—1985* (Scientific Report 68). Dhaka, Bangladesh: Author.
- Manton, K. G., Singer, B., & Woodbury, M. A. (1992). Some issues in the quantitative characterization of heterogeneous populations. In J. Trussell, R. Hankinson, & J. Tilton (Eds.), *Demographic application of event history analysis* (pp. 9–37). Oxford, UK: Clarendon.
- Moreno, L. (1994). Frailty selection in bivariate survival models: A cautionary note. *Mathematical Population Studies*, 4, 225–233.
- Trussell, J., & Rodríguez, G. (1990). Heterogeneity in demographic research. In J. Adams, D. A. Lam, A. I. Hermalin, & P. E. Smouse (Eds.), *Convergent issues in genetics and demography* (pp. 111–132). Oxford, UK: Oxford University Press.
- Vaupel, J. W. (1990). Relatives' risks: Frailty models of life history data. *Theoretical Population Biology*, 37, 220–234.
- Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439–454.
- Vaupel, J. W., & Yashin, A. I. (1985a). The deviant dynamics of death in heterogeneous populations. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 179–211). San Francisco: Jossey-Bass.
- Vaupel, J. W., & Yashin, A. I. (1985b). Heterogeneity's ruses: Some surprising effects of selection on population dynamics. *American Statistician*, 39, 176–185.
- Vermunt, J. K. (1997). *Log-linear models for event histories*. Thousand Oaks, CA: Sage.
- Wood, J. W., Holman, D. J., Yashin, A., Peterson, R. J., Weinstein, M., & Chang, M.-C. (1994). A multi-state model of fecundability and sterility. *Demography*, 31, 403–426.
- Wood, J. W., & Weinstein, M. (1990). Heterogeneity in fecundability: The effect of fetal loss. In J. Adams, D. A. Lam, A. I. Hermalin, & P. E. Smouse (Eds.), *Convergent issues in genetics and demography* (pp. 171–188). Oxford, UK: Oxford University Press.

UNOBTUSIVE MEASURES

Unobtrusive measures is the title of the classic reference for unobtrusive research procedures by Webb, Campbell, Schwartz, and Sechrest (1966; see also the revised edition by Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). An excellent summary has also been published by Lee (2000); see also UNOBTUSIVE METHODS. Instead of *measures*, the more inclusive term *unobtrusive methods* is now more frequently used because it includes methods other than the quantitative ones focused on by Webb et al.

“Unobtrusive measures” are research procedures designed to avoid a major problem in much social science research, namely, REACTIVITY (Webb et al., 1966). Research participants typically “react” to their situation; that is, they perceive or suspect that they are being observed or evaluated in some fashion. If one measures length with a ruler, the measurement process does not affect the accuracy or validity of the observation. If, however, attitudes about an issue are assessed using a questionnaire, survey, or interview procedure, then the data will likely be “impure,” that is, affected by participants' reactions and characteristics. For example, a QUESTIONNAIRE or SURVEY about racial tolerance or sexual matters will likely elicit participants' reactive concerns about the purpose of the research (e.g., who will see the data and what type of responses will seem appropriate). Data are also sensitive to factors such as the gender, race, age, or other characteristics of research personnel. Moreover, some research participants are motivated to “help” researchers, whereas others are less—sometimes, far less—cooperative. Although their hypotheses may sometimes be inaccurate, participants nevertheless usually respond in line with these speculations.

Frequently, researchers themselves have ideas or wishes about how their research will work out. We know that these expectations may affect how the work is carried out (Rosenthal, 1976), as well as subsequent results.

In contrast, unobtrusive methods are those in which participants are unaware of being observed or used as sources of data. The researcher(s) here believe it is important to avoid the usual requirement of obtaining participants' prior consent, that is, their active cooperation. Unobtrusive research often avoids human involvement altogether and makes use of historical or archival material or research strategies such as hidden

tape recorders, microphones, or even hidden observers. Other studies represent research on “say and do”; that is, they compare what participants say they would or would not do in certain situations with what they are actually observed to do. For example, one might telephone landlords advertising rooms or apartments for rent, make simple enquiries as to availability, and then compare the results with those from other conditions in which landlords are led to believe that callers are of a certain race or bear some stigmatizing characteristic, such as being homosexual, a person with AIDS, or a former psychiatric patient needing accommodation. Lee (2000) has also discussed the Internet as a domain for unobtrusive research, such as analysis of “home pages.”

Unobtrusive measures bear ethical pitfalls because their data do not assume the active cooperation or consent of participants. Yet a major goal of social science remains that of gathering, comparing, and contrasting observations from both obtrusive and unobtrusive measures.

—Stewart Page

REFERENCES

- Lee, R. (2000). *Unobtrusive methods in social research*. Buckingham, UK: Open University Press.
- Rosenthal, R. (1976). *Experimenter effects in behavioral research* (Rev. ed.). New York: Irvington.
- Webb, E., Campbell, D., Schwartz, R., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally.
- Webb, E., Campbell, D., Schwarz, R., Sechrest, L., & Grove, J. (1981). *Nonreactive measures in the social sciences*. Dallas: Houghton-Mifflin.

UNOBTUSIVE METHODS

For many years, social scientists have stressed the formal aspects of research—we thus have many textbooks on statistics and issues of RESEARCH DESIGN. One problem remains: We know much more about how our research is supposed to transpire than we do about how it actually does. In psychology, much research has assumed, for example, that human beings will simply write down or tell us verbally what they think, feel, or will do or that they will just “respond” automatically to research “stimuli.” Yet we often find a clear contrast between what people say (or what

they say they do) and what they actually do. A related contrast is that between what they do when they know they are being observed and what they do when they assume their behavior is private and unknown to, or unobserved by, others. *Unobtrusive methods* is the name given to research methods that neither depend on nor seek the active involvement or consent of human research participants and frequently are undertaken without participants’ awareness. Some methods, such as CONTENT ANALYSIS, gather data and observations from impersonal sources such as magazines, books, newspapers, Internet material, media presentations, or archival sources.

BACKGROUND

As an illustration of unobtrusive methods used to observe human behavior, the American television series *Candid Camera* became notorious but also respected by social scientists. The series was based on unobtrusively filmed records of people, observed without their knowledge in a variety of real social situations. These individuals proceeded to “be themselves”—usually to their later embarrassment. The staff of *Candid Camera*, for example, removed the engine from a car at the top of a large hill. The car was then “driven” without power down the hill and coasted directly into a service station at the bottom. When the (female) driver said she thought the engine was “running rough,” several (male) station attendants raised the hood of the car; peered in, then appeared stumped by the engine problem they saw, even though there was no engine there to see.

Unobtrusive means *noninterfering*, *silent*, or *in the background*. Thus, a simple though frequently forgotten lesson is taught by unobtrusive methods. That is, whenever appropriate, we should consider research methods that do not themselves adulterate or somehow affect or alter the nature of the data gathered. There are no specific types of quantitative or qualitative data analysis that uniquely define unobtrusive research methods. Moreover, there are neither limitations nor rigid boundaries for defining them. The term *unobtrusive* thus refers only to a general class of methods.

Unobtrusive methods developed historically from the increasing realization, since the 1960s, that many forms of social science research, particularly those involving social interaction between researchers and research participants, are subject to numerous sources

of unreliability and error. The common goal of these methods is to avoid what Webb, Campbell, Schwartz, and Sechrest (1966) referred to originally as REACTIVITY, that is, the tendency of many conventional (obtrusive) methods to give away or communicate to participants what the research is about—what they are looking for, as it were. In turn, participants then “react”; that is, they behave differently than would have been the case had their data been gathered unobtrusively and without their awareness of being a focus of research. One form of reactivity is that the participants perceive or guess correctly what the research objectives or hypotheses are, basing this on cues detected during data gathering. Moreover, we know that many respondents, for example, will respond to QUESTIONNAIRE or SURVEY items to create a “good impression” or respond according to other forms of bias. In some cases, participants’ guesses about a research study may be uncertain or inaccurate, but they will nevertheless affect their responses and behavior. There are multiple sources of data reactivity, including not only participants’ perceptions and hypotheses about the nature and purpose of the research but also effects exerted by characteristics of the researcher and research situation itself.

In recognition of this issue, Webb et al. (1966; see also the revised edition by Webb, Campbell, Schwartz, Sechrest, & Grove, 1981) published their volume *Unobtrusive Measures*. The field of social science owes a great debt to these authors because our knowledge and awareness of these procedures trace largely to their writings. A valuable summary of this area, with many recent examples, has also been provided by Lee (2000). Although the point is often missed, Webb et al. emphasized that they did not totally object to conventional research methods. They indicated, however, that these measures, sometimes of necessity, were often used naively and alone—that is, without being supplemented by or subject to verification from unobtrusive methods examining the same research question or issue. Given this situation, social science often has difficulty in establishing the boundaries of validity and generalizability for its findings. We too seldom know, for example, how well conventional testing or data-gathering procedures will equip us to predict behavior in real-life (as opposed to experimental or laboratory) situations. Conventional methods, as referred to above, are familiar: interviews, psychological tests, qualitative studies, rating scales, surveys, and a host of one-shot studies based on simple, often quite obvious

or transparent questionnaire measures. Thus, in social science, we frequently find, rhetorically put, that the presence and characteristics of our microscope actually serve to cause changes in the behavior of the material or organism under study.

EXAMPLES OF UNOBTUSIVE METHODS

Thus, we describe the unobtrusive or nonreactive method as one that seeks to avoid reactivity and does not require or seek the active cooperation, awareness, or consent of participants involved. Many such studies might therefore be called “say and do” research or, more accurately, say *versus* do. Indeed, Lee (2000) classifies many studies of this type as FIELD EXPERIMENTS. We list a few examples below, for which relevant references and further details may be found in Page (2000). Thus, we might assess the following:

1. The popularity of library or museum materials by unobtrusively measuring and subsequently comparing the extent of disintegration, per unit of time, on the floor surfaces surrounding different types of exhibits (Webb et al., 1966).

2. The extent of customers’ interest in buying certain products, as indicated by extent of pupil dilation (Webb et al., 1966).

3. The extent of racial prejudice in a given locality by observing the extent to which landlords advertising rooms or apartments for rent will communicate false information about room availability to individuals believed to belong (or not belong) to a certain race or to bear some type of stigmatizing characteristic such as being homosexual, a person with AIDS, or a former psychiatric patient now needing accommodation. A variation of this basic method has been used in civil rights and desegregation efforts, that is, in which rooms known to be unrented and available are enquired about by Black persons as potential tenants, sometimes making personal visits to landlords to gather this information.

4. The extent to which clinical therapists or distress center counselors show empathy and related helping skills with their clients by analyzing unobtrusively obtained (without participants’ knowledge) tape recordings of online, telephone-based, or live interactions.

5. The extent to which respondents will indicate high acceptance and tolerance toward Blacks, for example, as determined from an interview,

psychological test, or survey procedure yet be less likely to patronize Black as opposed to White cashiers in cafeterias or to show subtle (e.g., nonverbal or gestural) forms of rejection or hostility when these behaviors are observed unobtrusively over time.

6. The extent to which potential community employers will offer or deny employment to persons alleging to have been former psychiatric patients compared to other individuals making no such reference. Several studies using this basic approach, sometimes including the use of hidden tape recorders, have been performed by Professor Amerigo Farina and colleagues at the University of Connecticut.

7. The extent to which mail, believed to be written to (or from) a psychiatric patient and accidentally "lost" in a variety of locations, will be forwarded (posted) by persons finding such mail.

8. The extent to which differences will be found when students' evaluations of their courses or professors, based on interviews or mandatory written assessments, are compared with the content of unobtrusively obtained recordings of students' waiting-room or cafeteria conversations.

9. The extent to which Black shoppers, unobtrusively observed, might receive poorer service in stores compared to service to Whites or be quoted higher prices for used cars or other merchandise compared to those given to Whites.

10. The extent to which changes in students' morale and school spirit at a university, over time, are reflected in changes in bookstore sales of clothing and supplies bearing their university's colors, logo, or name.

11. The nature of a society's attitudes toward the mentally ill, including changes in these over time, by examining the content of newspaper articles and news items referring in some fashion to the label of "mental illness." One might conduct research of this nature, for example, by examining all Canadian newspaper items and stories archived under this heading in the Canadian Newspaper Index and then comparing their content over both recent and past time periods. The use of archival or "running" records (Lee, 2000) or CONTENT ANALYSIS (i.e., obtaining observations from already existing sources)—quite apart from direct involvement or observation of human participants—is now a popular form of unobtrusive data collection.

A final example concerns the increasing attractiveness of using the Internet to gather unobtrusively obtained data and observations. The Internet

represents a data source that enables the unobtrusive study to explore both the content and methods of self-expression in which online users engage. In addition, the researcher in this domain becomes largely freed of the usual research constraints of time, space, and cost. An excellent discussion of Internet-based research, addressing these points and with many examples, may be found in Lee (2000). For example, the style and content of "home pages" may be studied as a means of learning about individuals, groups, or organizations. Sometimes, information about sensitive research topics may be gathered if one can establish reliable and confidential online contacts. The anonymous nature of much Internet "behavior" also affords an opportunity to study online users in a virtual environment existing quite apart from normal concerns about how one will "look." Patterns of computer usage, communications, and online postings from "chat rooms," online discussions, or newsgroups may also be studied.

ETHICAL CONSIDERATIONS AND CURRENT STATUS OF UNOBTUSIVE METHODS

Moreover, assessing the current status of unobtrusive methods includes considerations about the issue of obtaining INFORMED CONSENT from participants. First we note that many common research methods, prevalent at the time of the first Webb et al. (1966) book, remain still quite popular and prevalent today. Perhaps this reflects the issue of consent; that is, they remain in vogue not necessarily through any inherent or absolute guarantee of validity but because it is usually not difficult to obtain and document participants' consent to take part in them. This perspective is congruent with the current popularity of SOCIAL CONSTRUCTIONISM, that is, the view that research participants must be viewed as genuinely participating in research for which they are providing data. The constructionist view thus sees participants as helping to create the results, as having a "research relationship," a type of coexistence, with the researchers. In this view, facts and knowledge are not discovered (as we generally assume they are) but instead are invented or jointly created by social consensus. (Research participants in psychology, until recently, were called *subjects*, a term that, for social constructionists, implied that they were inert and object-like rather than human beings helping researchers to create knowledge.)

Consequently, then, unobtrusive methods, based generally on data gathering without the cooperation or knowledge of research participants, are hardly in vogue today in terms of methodology or epistemology of research. Of course, for Webb et al. (1966), the active involvement, consent, and cooperation of participants, although appropriate in many contexts, were not intrinsic goals of research. To the contrary, they represented sources of reactivity to be avoided or at least supplemented by other measures. In essence, the dependence on and, in fact, the seeking of data from actively consenting (and collaborating) research participants are effectively a rejection of the very purpose and goals of unobtrusive measures (Page, 2000). Unobtrusive methods are seldom emphasized in current research courses or in the research work of students, despite our concerns about the generalizability and validity of data from reactive methods. Webb et al. reported in the revised (1981) edition of their earlier book, as cited by Lee (2000), that although most social scientists seem generally to have recognized the value of unobtrusive measures, the actual use of the methods is now infrequent.

Unobtrusive measures do involve ethical pitfalls, and in some cases, their use may even be legally precarious. It may be problematic or awkward to record or observe certain behaviors without participants' knowledge or, as some have done, to create simulated emergency situations in public places to observe reactions of the public. Data from unobtrusive methods may sometimes be unreliable or difficult to interpret, no less so than with other methods, so that in some instances, one must consider them with some trepidation. Another question is whether a study should refrain from using unobtrusive measures when another method might be used in which informed consent could be obtained from participants, that is, if the alternative method will not be unduly affected by reactivity. Lee (2000) also describes certain ethical issues arising from unobtrusive studies using the Internet. Many of these concern the difficulty in drawing the line between public and private domains and in safeguarding the confidentiality of online participants where such is critical. We note here that Webb et al. (1966) did not view nonreactive methods as inherently unethical and, in fact, saw them as enabling researchers to sleep better, that is, to gather data without inconveniencing, concerning, or otherwise interfering with the lives of participants.

Increased consideration of unobtrusive methods, consistent with the work of Webb et al. (1966, 1981)

and others, has the potential to augment conventional methods and to move us beyond our dependence on reactive data—thereby effectively reacquainting us with the basic issue of the generalizability of research observations as well as the need to “triangulate,” that is, to use different types of measures in researching a particular problem or hypothesis. We also note Donald Campbell's belief that social science research should ultimately be conducted for others, that is, for the benefit of the community and society. Research using some form of nonreactive methodology, sometimes involving some form of DECEPTION of participants where this can be scientifically justified, has the potential to generate results that are striking, unexpected, and often counterintuitive—and therefore of particular interest and value.

—Stewart Page

REFERENCES

- Lee, R. (2000). *Unobtrusive methods in social research*. Buckingham, UK: Open University Press.
- Page, S. (2000). The lost art of unobtrusive measures. *Journal of Applied Social Psychology*, 30(10), 2126–2128.
- Webb, E., Campbell, D., Schwartz, R., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand-McNally.
- Webb, E., Campbell, D., Schwartz, R., Sechrest, L., & Grove, J. (1981). *Nonreactive measures in the social sciences*. Dallas: Houghton-Mifflin.

UNSTANDARDIZED

The concept of *unstandardized* can best be understood in contrast to the term *standardized*. Standardization transforms statistics of different metrics to a common metric to permit comparison. In other words, if the statistics are unstandardized (have different terms of measurement), they cannot be compared. Take, for example, the PARTIAL UNSTANDARDIZED REGRESSION COEFFICIENT, or the slope, of a multiple regression equation. Consider a time series of data about Brazil, 1970–1991 (World Bank, 1997):

$$\hat{Y} = -5764 + 110X_1 + 28X_2,$$

where $\hat{Y}$ is the gross national product (GNP) per capita in U.S. dollars, X_1 is the percentage urban population, and X_2 is the net capital inflow in billions of U.S. dollars. Both variables, X_1 and X_2 , turn out to be

significant. The results indicate that a 1% increase in the percentage urban population of Brazil is, on average, associated with an increase in per capita GNP of \$110, and a \$1 billion increase in net capital inflow to Brazil is, on average, associated with an increase in per capita GNP of \$28. However, the unstandardized partial regression coefficients, or slopes, do not indicate which variable, urban population or net capital inflow, has a greater impact on the dependent variable per capita GNP. Although the value of the slope coefficient is greater for the variable percentage urban population, clearly a 1% increase in urban population is qualitatively different from an increase of \$1 billion in net capital inflow. In fact, I can artificially increase the slope coefficient (unstandardized partial regression coefficient) for net capital inflow by changing the terms of measurement to trillions of dollars from billions of dollars. The sizes of the slopes, the unstandardized regression coefficients, do not indicate which independent variable is associated with a greater change in the dependent variable.

To facilitate comparison of the impact of the INDEPENDENT VARIABLES on the DEPENDENT VARIABLE, we can standardize the variables on a common scale to produce STANDARDIZED REGRESSION COEFFICIENTS, also known as beta weights. The values of the independent and dependent variables are standardized by calculating their position in terms of STANDARD DEVIATIONS away from the mean. This procedure produces new values for the variables that are then used to reestimate the regression equation (this procedure can be performed by statistical packages when analyzing the data). The beta weight, or the standardized regression coefficient, is “the average standard deviation change in Y associated with a one standard deviation change in X , when the other independent variables are held constant” (Lewis-Beck, 1980, p. 65). This transformation now permits assessment of which independent variable is associated with a greater change in Y . In the example here, we find that a 1 standard deviation change in urban population is associated with a .99 standard deviation change in GNP per capita, whereas a 1 standard deviation change in net capital inflow is associated with a .17 standard deviation change in GNP per capita. Clearly, urban population seems to be more important for explaining changes in GNP per capita, but net capital inflow is of marginal importance.

The standardization procedure takes advantage of the known properties of the NORMAL DISTRIBUTION to permit comparison of independent variables

in a model. Standardization of partial regression coefficients, however, only permits comparison within a sample. Beta weights cannot be compared across samples to determine, for example, whether the percentage urban population is similarly associated with a large change in per capita GNP in Argentina, assuming the VARIANCES of the independent and dependent variables differ between Brazil and Argentina. A comparison of beta weights across samples can thus artificially alter the value of the beta weights and lead to possibly erroneous conclusions. Instead, it is appropriate to compare a variable, such as the effect of urban population on per capita GNP, across samples with the unstandardized regression coefficient because it is not sensitive to different X variances, and the variables are of the same metric in both samples.

The concept of standardization is employed in other statistical measures and concepts such as the STANDARD NORMAL DISTRIBUTION, STANDARD SCORES, and the CORRELATION COEFFICIENT. The correlation coefficient is a measure of the strength of ASSOCIATION between two variables. This measure is not susceptible to changes in the terms of measurement that afflict the measure of COVARIANCE because the correlation coefficient simply standardizes the covariance by using the standard deviations of each variable to create standard scores. Standardization of the covariance renders a statistic of ASSOCIATION that is more intuitive and thus more useful than the covariance statistic. Because the covariance statistic lacks an upper or lower bound, it does not provide a measure of the strength of the association. We do not know whether the magnitude of the covariance value reflects a strong association or is a product of the terms of measurement of the variables. Standardization ameliorates this problem by forcing the measure of association within the interval of -1 and 1 , with -1 or 1 representing perfect linear association.

The standard normal distribution and standard scores provide the foundation on which much of statistical analysis is executed. The examples of the beta weight and the correlation coefficient illustrated here both use these concepts. Standard scores (also known as Z -scores or standardized variables) are obtained by standardizing the variables using the concept of a standard normal distribution. A standard score of a variable X is expressed as the sample mean subtracted from the values of the variable x_i and then divided by the standard deviation s . The result is a variable with a standard normal distribution—a mean of zero

and a variance of 1. As these examples illustrate, the essence of standardization is to find a common metric on which statistics or variables of different metrics can be evaluated.

—Jacque L. Amoureux

REFERENCES

- Lewis-Beck, M. S. (1980). *Applied regression: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–022). Beverly Hills, CA: Sage.
- Lewis-Beck, M. S. (1995). *Data analysis: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–103). Thousand Oaks, CA: Sage.
- World Bank, International Economics Department. (1997). *World tables of economic and social indicators, 1950–1992* [Computer file]. Ann Arbor, MI: Interuniversity Consortium for Political and Social Science Research.

See also STANDARDIZED REGRESSION COEFFICIENT

UNSTRUCTURED INTERVIEW

The unstructured interview comprises various types: creative, ACTIVE, postmodern, and feminist interviews. They all share a common concern with allowing interviewees as much latitude as possible in answering OPEN-ENDED QUESTIONS and going off in directions of their own. The interview is seen as an interaction accomplished between the interviewer and the interviewee.

The unstructured interview, also called an in-depth interview, does not rely on CLOSED-ENDED or structured questions. The interviewer pursues information about a given topic by asking open-ended questions or merely prompting the interviewee. The interview may be taped or the interviewer may take notes during the interview or, more rarely, shortly after the conclusion of the interview.

For the interview to take place successfully, the interviewer must gain the trust of the respondent. The interviewer may gain trust by relying on his or her academic status and the scope of the study or by temporarily becoming a member of the group under observation; at times, the respondents are paid to participate in the interview (cf. Malinowski, 1989). Unstructured interviews are often practiced in conjunction with PARTICIPANT OBSERVATION; in fact, many of the Chicago school classic studies of the

1930s relied almost entirely on unstructured interviews (Harvey, 1987).

Unstructured interviews have recently taken many forms that differ somewhat from each other. They can be divided into the following subgroups:

Creative Interviewing. In *Creative Interviewing*, Jack Douglas (1985) proposes different techniques for unstructured interviewing, based on his research and that of his graduate students. Traditionally, the interviewer attempts to remain neutral and not to influence the study in any way. Thus, it should not matter who the interviewer is *qua* individual. Douglas feels differently. He argues that there is an unwarranted assumption in unstructured interviewing that the respondent will be truthful in his or her answers. Yet, muses Douglas, people lie all the time in everyday life, so why should respondents be any different? To maximize the chances of truthfulness and full disclosure, Douglas feels that the interviewer needs to offer a *quid pro quo*. The interviewer must disclose a private glimpse in his or her private life to the respondent, thus becoming truly a human being rather than an impersonal neutral interviewer and gaining, it is hoped, the trust of the respondent. In addition, Douglas indicates that information obtained by respondents must be continually checked for veracity by the use of informants. Informants tend to be marginal members of the group studied, and hence they tend to be discontented and likely to divulge information about their group.

ACTIVE INTERVIEW. James Holstein and Jaber Gubrium (1995) moved away from interviews as a one-sided affair in which the interviewer maximizes techniques to gather the best amount of information from the respondent. Instead, Holstein and Gubrium consider the unstructured interview as an active interaction between the interviewer and the respondent. The data gathered are far from neutral. Rather, the interview is the negotiated result reached together by the two (or more) individuals. The results depend on the level of collaboration and ability to interact by the two, who together create a story—the interview. The interview does not take place in a vacuum but is affected by the context in which it occurs. An added relevant dimension for the interview, for Holstein and Gubrium, is determined by *how* the results are achieved (the reflexive analysis of the elements that affect the interview), as well as by *what* is achieved (the substantive findings of the interview).

Postmodern Interview. The postmodern interview is another subgroup of the unstructured interview. POSTMODERNISM is a philosophical set of notions that has permeated many disciplines. Although highly controversial, it has brought some important considerations to the unstructured interview. Postmodernism advocates the demise of meta-theories and PARADIGMS and wishes to revert to studying and understanding the minutiae of everyday life, as viewed by the members of society themselves. Clearly, the unstructured interview can be informed by postmodern notions, especially when addressing the concern with the power structure of the unstructured interview. Traditionally, the interviewer controls the interview—asks the questions, determines what is relevant, decides on the duration of the interview, and chooses how to use the results of the interview. Postmodernists feel that the respondents are being exploited, and to use them and their information for purposes that do not benefit them at all is unethical. Postmodern sociologists reject this model of interviewing in favor of an overtly politically cooperative one. They advocate the study of unprivileged groups, with the open agenda of helping them. This flies in the face of the traditional sociological notion of neutrality and OBJECTIVITY. What results is open partisanship whereby the results of the interview are to be used in a way to aid the group being studied.

Feminist Interview. A final subgroup comes from the feminist interview. Often, the feminist interview is coupled with the longstanding ORAL HISTORY approach. Feminists consider traditional unstructured interviews as highly paternalistic and biased. Both the interviewers and the concerns of the studies tend to be overwhelmingly male centered. For feminists, it is time to be attentive to the concerns of the women interviewed. Also, by introducing female interviewers, the tension in the interview is greatly reduced. Woman-to-woman interviews open the door to more open disclosure of information, as the respondents are freed from the subordinate roles they take in traditional male-female interviews. Thus, they may feel free to address problems and issues they had previously felt unable to disclose. Feminist interviews are intended to help the women studied, not use them and their information.

In concluding, keeping all of the aforementioned variations in mind, the unstructured interview remains a fundamental tool in the quest for sociological

knowledge. The methodology of asking questions and getting answers allows sociologists to understand the meaning of everyday life for its members. The unstructured nature of the interview does not foreclose on the possible answers but leaves a wide-open field, fostering a broader understanding of society.

—Andrea Fontana

REFERENCES

- Douglas, J. D. (1985). *Creative interviewing*. Beverly Hills, CA: Sage.
- Harvey, L. (1987). *Myths of the Chicago school of sociology*. Aldershot, England: Avebury.
- Holstein, J., & Gubrium, J. (1995). *The active interview*. Thousand Oaks, CA: Sage.
- Malinowski, B. (1989). *A diary in the strict sense of the term*. Stanford, CA: Stanford University Press.

UTILIZATION-FOCUSED EVALUATION

Utilization-focused evaluation premises that EVALUATIONS should be judged by their utility and actual use; therefore, evaluators should facilitate the evaluation process and design any evaluation with careful consideration of how everything that is done, *from beginning to end*, will affect use. The *focus* is on *intended use by intended users*. The evaluator develops a working relationship with intended users to help them determine what kind of evaluation they need. Utilization-focused evaluation does not advocate any particular evaluation content, model, method, theory, or even use. Rather, it is a process for helping primary intended users to select the most appropriate content, model, methods, theory, and uses for their particular situation. The evaluator collaborates with an identified group of primary users, focusing on their intended uses of evaluation.

—Michael Quinn Patton

REFERENCE

- Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

V

V. See SYMMETRIC MEASURES

VALIDITY

One of the most important steps in social research is the MEASUREMENT of theoretical CONSTRUCTS. Relating the abstract CONCEPTS described in a THEORY to empirical INDICATORS of those concepts is crucial to being able to empirically examine and test the phenomena under study. The association of an abstract theoretical concept with its empirical manifestation is at the heart of validity.

Validity is generally defined as the extent to which any measuring instrument measures what it is intended to measure. Strictly speaking, however, the only way a researcher can access the validity of a measurement is by validating “an interpretation of data rising from a specified procedure” (Cronbach, 1971, p. 447). The clear implication here is that a measuring instrument can be relatively valid for measuring one phenomenon but invalid for measuring other phenomena. Thus, the measuring instrument itself is not validated, but the measuring instrument in relation to the purpose for which it is being used.

There are three basic types of validity that concern us here: CRITERION-RELATED VALIDITY, content validity, and CONSTRUCT VALIDITY. Each type evaluates validity in a different way, and each has its own uses and limitations.

Criterion-related validity, or predictive validity, is closest to the definition above. A measure is said to be relatively valid if it accurately predicts the results on some other, external measure, or criterion. For example, one validates a written airline pilot’s examination by showing that it accurately predicts how well a group of people can actually fly an airplane. The degree of criterion-related validity is indicated by the correspondence between the test and the criterion. Indeed, this CORRELATION is the only kind of evidence that is relevant to criterion-related validity.

There are two specific types of criterion-related validity. When the criterion exists in the present and is correlated with a current measure, we are dealing with concurrent validity. Predictive validity, on the other hand, focuses on a situation in which the measure predicts a future criterion or outcome. The difference between these two types of criterion-related validity thus concerns the temporal placement of the criterion—now or in the future.

Content validity is another basic type of validity. This type focuses on the extent to which a particular empirical measurement reflects a specific domain of content. For example, a test in world history would not be content valid if it focused only on the history of the United States or Europe. A researcher must take into account several considerations when assessing content validity. First, the researcher must be able to specify the full domain of relevant content. In the example above, a world history test would need to cover all regions of the world as well as the entire period of recorded history. For many abstract theoretical concepts found in the social sciences, it often proves difficult to specify the full domain of content. Second, the researcher must

take a careful sample of this domain for testing. The difficulty here is that it is not always a sample from the full domain of content. Third, the sample of content must be put into a form suitable for testing. Thus, while intuitively appealing, content validation is often difficult to assess in the social sciences.

The third basic type of validity is construct validity. Fundamentally, construct validity focuses on the extent to which a measure performs according to theoretical expectations. Assessing the construct validity of a measure involves three distinct but interrelated steps. First, one must specify the theoretical relationship between two or more theoretical constructs. Second, the empirical relationship between the measures of the constructs must be examined. And finally, the empirical relationship evidence must be interpreted in terms of how it clarifies the construct validity of the particular measure. Thus, if, on the basis of theoretical considerations, it is expected that racial prejudice is more prevalent among less educated people than among those with higher levels of education, one could assess the construct validity of a measure of racism by seeing if it is negatively correlated with education. This example shows that construct validation can be assessed only within a given theoretical context, because it is the context that generates the theoretical predictions that are at the heart of this type of validity. Because construct validation is woven directly into the fabric of social science theory, it is considered more pertinent in the social sciences than either criterion-related validity or content validity.

—Edward G. Carmines and James Woods

REFERENCES

- Carmines, E. G., & McIver, J. P. (1981). *Reliability and validity assessment* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-017). Newbury Park, CA: Sage.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Spector, P. E. (1992). *Summated rating scale construction: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-082). Newbury Park, CA: Sage.

VARIABLE

A variable is something that varies in value, as opposed to a constant (such as the number 2), which

always has the same value. These are observable features of something that can take on several different values or can be put into several discrete categories. For example, students' scores on an exam are variables because they have different values, and religion can be considered a variable because there are multiple categories. Scientists are sometimes interested in determining the values of constants, such as π , the ratio of the area of a circle to its squared radius. However, statistics involves the study of variables rather than constants.

A quantity X is a random variable if, for every number a , there is a probability p that X is less than or equal to a . A discrete random variable is one that attains only certain values, such as the number of newly elected senators in a given election. By contrast, a continuous random variable is one that can take on any value within a range, such as a person's height (measured in the smallest possible fractions of an inch).

Data analysis often involves HYPOTHESES regarding the relationships between variables, such as "If X increases in value, then Y tends to increase (or decrease) in value." Such hypotheses involve relationships between latent variables, which are abstract concepts. These concepts have to be operationalized into manifest variables that can be measured in actual research.

A basic distinction in statistical analysis is between the DEPENDENT VARIABLE that the researcher is trying to explain and the INDEPENDENT VARIABLES that serve as predictors of the dependent variable. In REGRESSION, for example, the dependent variable is the Y VARIABLE on the left-hand side of the regression equation $Y = a + bX$, whereas X is an independent variable on the right-hand side of the equation.

The starting point in statistical analysis is often looking at the distribution of the variables of interest, one at a time, including calculating appropriate univariate statistics. The changes in that variable over time can then be examined in TIME-SERIES ANALYSIS. Univariate analysis is usually just the jumping-off point for BIVARIATE or MULTIVARIATE ANALYSIS. For example, in ANALYSIS OF VARIANCE, the researcher examines the extent to which experimental conditions affect the variance in the dependent variable.

—Herbert F. Weisberg

REFERENCE

- Lewis-Beck, M. S. (1995). *Data analysis: An introduction*. Thousand Oaks, CA: Sage.

VARIABLE PARAMETER MODELS

It is usually assumed that the parameter, BETA (β), in a REGRESSION equation relating the INDEPENDENT and DEPENDENT VARIABLES is constant over all observations. But for TIME SERIES, it is possible to model the evolution of the β parameter over time, usually assuming that this evolution follows some common continuous time series process. Thus, the simplest time series variable parameter model is

$$Y_t = \beta_t X_t + \varepsilon_t,$$

where $\beta_t = \rho\beta_{t-1} + v_t$.

Such a model can be estimated by MAXIMUM LIKELIHOOD ESTIMATION methods via the Kalman filter. It is also possible to test for whether the parameters do vary over time (either continuously, as above, or more discretely). These models are also known as stochastic parameter regression models.

—Michael S. Lewis-Beck

See also RANDOM-COEFFICIENT MODEL

REFERENCE

Newbold, P., & Bos, T. (1985). *Stochastic parameter regression models* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 51-001). Beverly Hills, CA: Sage.

VARIANCE. See DISPERSION

VARIANCE COMPONENTS MODELS

Statistical models such as REGRESSION or the ANALYSIS OF VARIANCE partition observed variation into a systematic portion linked to explanatory variables or experimental treatments, and a random portion. Random variation is commonly represented by a single error term. When two or more sources of random variation affect a phenomenon, the random portion of variation can be represented using a variance components model.

Important social science applications of variance components models are in analyses of data from MULTISTAGE SAMPLING designs, generalizability studies assessing the quality of measurements (see GENERALIZABILITY THEORY), experiments including random factors, and longitudinal studies with repeated measures. Variance components models constitute the random part of MULTILEVEL ANALYSIS or HIERARCHICAL LINEAR MODELS.

Consider a simple model for an observation y_{ij} from a two-stage sample: $y_{ij} = \mu_y + a_i + e_{ij}$; μ_y is the population mean, a_i is a first-stage error term, and e_{ij} is a second-stage error term. The first-stage error terms a_i are known as random effects. In an educational study, the first-stage sampling units could be schools, the second-stage sampling units students within schools, and measure y_{ij} the mathematics achievement of the j th student sampled in the i th school. Here, one source of random variation (students) is nested within the other (schools). The corresponding variance components model is $\text{Var}(y_{ij}) = \sigma_a^2 + \sigma_e^2$, where σ_a^2 is the variance component for the first-stage sampling units (schools) and σ_e^2 is the variance component for the second-stage sampling units (students within schools).

The variance components measure the portions of random variation in y_{ij} attributable to different sources. For the example, the quantity

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

gives the proportion of total random variation assigned to the first-stage units (schools). Known as the INTRAClass CORRELATION, ρ also measures the degree of homogeneity among second-stage units clustered within a first-stage unit.

Using a variance components model takes interdependencies among observations into account when drawing inferences. In the example, one might seek to estimate μ_y from a sample consisting of n_i students from each of M schools, or a total of $N = \sum_{i=1}^M n_i$ observations. The n_i observations on students within a school are not independent, because they share a value of the school-level error a_i . Analyzing the sample as if it contained N independent observations would result in too small a standard error for the estimate of μ_y ; hence, confidence intervals for μ_y would be too narrow and p values associated with hypothesis tests about μ_y too low.

Estimation of variance components usually proceeds under assumptions that random effects are independently and identically distributed with means of 0. Analysis of variance (ANOVA) methods estimate variance components as functions of observed mean squares, making use of the fact that expected mean squares in the analysis of variance can be expressed as linear combinations of variance components. MAXIMUM LIKELIHOOD (ML) and restricted maximum likelihood (REML) methods obtain estimates for variance components by maximizing a likelihood function. This requires assumptions about the distributional forms of random effects, ordinarily specifying normality. ML and REML have become the preferred criteria for estimating variance components. These estimates usually must be obtained via iterative techniques.

As an empirical example, the model above was applied to scores on a 10-item vocabulary test administered to 1,015 respondents nested within 111 primary sampling units (PSUs) in the 1993 General Social Survey. REML parameter estimates are $\hat{\mu}_y = 6.00$, $\hat{\sigma}_a^2 = 0.44$, and $\hat{\sigma}_e^2 = 4.09$. The total variance is 4.53, and the estimated intraclass correlation is 0.10: About 10% of the variance in vocabulary scores lies between PSUs, with the remaining 90% within PSUs. The estimated standard error of $\hat{\mu}_y$ is 0.09, so a 95% confidence interval for the mean is (5.82, 6.18). Ignoring the clustering of respondents within PSUs and analyzing the 1,015 observations as if they were independent, the standard error of the mean is 0.07 and the confidence interval for μ_y correspondingly shorter.

Studies with more than two sources of random variation require more elaborate variance components models. These take different forms depending on the number of random sources and how they are crossed with or nested within one another. In a generalizability study about job performance ratings, for example, several raters might judge performance quality for each of several employees with respect to several job tasks. Raters, tasks, and employees are crossed sources of random variation in the ratings. Assuming no interactions among the three sources, a model for the rating y_{ijk} of the i th employee by the j th rater with respect to the k th task is $y_{ijk} = \mu_y + a_i + b_j + c_k + e_{ijk}$, where μ_y is the mean rating and a_i , b_j , and c_k are the random effects of employees, raters, and tasks, respectively. The variance $\text{Var}(y_{ijk})$ is then modeled as $\sigma_a^2 + \sigma_b^2 +$

$\sigma_c^2 + \sigma_e^2$, where the four variance components refer to employees, raters, tasks, and error, respectively. These variance components are used in calculating generalizability coefficients (analogous to intraclass correlation coefficients) reflecting the dependability of the ratings.

In multilevel regression models, random variation in parameters leads to other variance component models. An example is the simple linear regression of student j 's mathematics achievement (y_{ij}) on IQ (x_{ij}) within the i th of a sample of schools: $y_{ij} = a_i + b_i x_{ij} + e_{ij}$. The school-specific intercepts a_i vary randomly around an average intercept a as follows: $a_i = a + u_i$, whereas the school-specific slopes b_i vary randomly around an average slope b : $b_i = b + v_i$. The random effects u_i and v_i on the parameters of the school-level regressions are sources of random variation in scores y_{ij} , which may be rewritten as $y_{ij} = a + b x_{ij} + u_i + v_i x_{ij} + e_{ij}$. The variance of achievement given IQ is then $\text{Var}(y_{ij}|x_{ij}) = \sigma_u^2 + x_{ij}^2 \sigma_v^2 + 2x_{ij} \sigma_{uv} + \sigma_e^2$, where σ_u^2 is a school-level variance component for intercepts, σ_v^2 is a school-level variance component for slopes, σ_e^2 is a variance component for students within schools, and σ_{uv} is a covariance between the school-level random effects on intercepts and slopes. Quite complex structures of random variation can arise in models with several levels of nesting and multiple random coefficients.

—Peter V. Marsden

REFERENCES

- Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

VARIANCE INFLATION FACTORS

Variance inflation factors (VIFs) measure the impact of COLLINEARITY among the predictors in a REGRESSION analysis on the precision of ESTIMATION. The “variances” in question are the SAMPLING variances of the REGRESSION COEFFICIENTS. The term *variance inflation factor* was possibly coined by Donald Marquardt (1970).

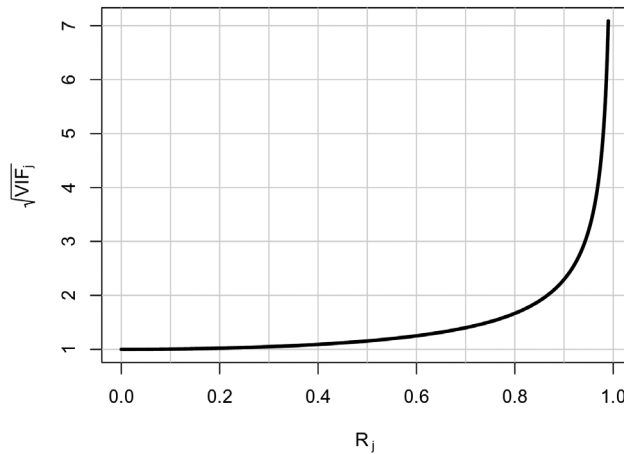


Figure 1 The Square Root of the Variance Inflation Factor VIF_j as a Function of the Multiple Correlation R_j From the Regression of x_j on the Other Predictors

Consider the LINEAR REGRESSION model

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i,$$

where y_i is the value of the response variable for the i th of n observations; $x_{1i}, x_{2i}, \dots, x_{ki}$ are the PREDICTOR VARIABLES; ε_i is the error, usually assumed to be distributed independently of the errors for the other observations, with zero mean and constant variance σ^2 ; α is the regression constant; and (in a model in which the predictors are quantitative) the β s are slope coefficients.

The sampling variance of the LEAST-SQUARES estimate b_j of β_j is

$$V(b_j) = \left[\frac{\sigma^2}{(n-1)s_j^2} \right] \left[\frac{1}{1-R_j^2} \right], \quad (1)$$

where $s_j^2 = \sum (x_{ji} - \bar{x})^2 / (n-1)$ is the variance of the j th predictor, and R_j^2 is the squared multiple correlation from the regression of x_j on the other x s. The second factor on the right-hand side of equation (1), $1/(1-R_j^2)$, called the variance inflation factor for b_j , or VIF_j , expresses the degree to which collinearity among the predictors degrades the precision of b_j , relative to similarly dispersed uncorrelated predictors. The square root of the variance inflation factor expresses the impact of collinearity on the size of the confidence interval for β_j .

It is worth mentioning the other, and more common, sources of imprecision in estimation revealed by equation (1): large error variance, σ^2 ; small sample size, n ; and predictors with small dispersion, s_j^2 . As well, it should be pointed out that collinearity must be very strong before the precision of the regression estimates is substantially degraded. Figure 1 plots the square root of the variance inflation factor against the multiple correlation R_j ; even when R_j is as large as 0.8, for example, the square root of the VIF is still less than 2. Multiple correlations among predictors as large as 0.8 are uncommon in social science data.

Variance inflation factors are applicable to one-coefficient terms in linear models, such as quantitative predictors with linear effects. Multiple-coefficient terms—such as a set of dummy regressors representing a categorical predictor—require a more general approach, because correlations among regressors in a related set are affected by inessential changes to the model (such as a change in baseline category for a set of dummy regressors). John Fox and Georges Monette (1992) suggest a generalization of variance inflation factors to cover these situations, comparing the relative sizes of the joint confidence region for a subset of coefficients for correlated and uncorrelated predictors.

The joint confidence region for two regression coefficients is bounded by an ellipse; for three coefficients, it is ellipsoidal, and for more than three coefficients, it is hyper-ellipsoidal. Figure 2(a) shows two “data ellipses”: The solid ellipse is for uncorrelated predictors x_1 and x_2 ; the axes of the ellipse are parallel to the x_1 and x_2 axes. The broken ellipse is for predictors that are highly correlated, $r_{12} = 0.95$; the tilt of the ellipse reflects the positive correlation between the predictors. Both ellipses are scaled so that their “shadows” on the x_1 and x_2 axes are the standard deviations of the predictors; note that both ellipses are centered at the means of the predictors. For bivariate-normal predictors, the data ellipse is also a contour of constant probability density, but regardless of the joint distribution of the x s, the data ellipse is a graphic representation of the dispersion and correlation of the predictors.

The ellipses in Figure 2(b) give 95% confidence regions for the regression parameters β_1 and β_2 assuming identical sample regression coefficients b_1 and b_2 ; these confidence ellipses are rescaled 90-degree rotations of the corresponding data ellipses. Notice that the

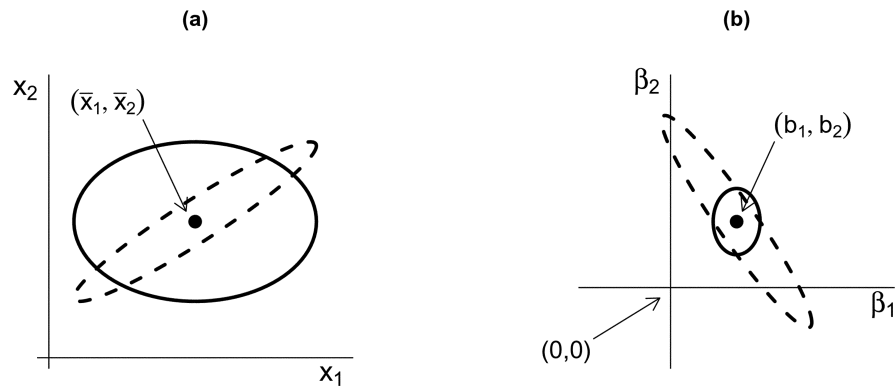


Figure 2 (a) Data Ellipses for Uncorrelated (Solid) and Highly Correlated (Broken) Predictors; (b) Corresponding 95% Confidence Ellipses

size of the confidence ellipse for the highly correlated predictors is much larger than the ellipse for the uncorrelated predictors, and that the former ellipse includes values of 0 for both β_1 and β_2 individually (although not simultaneously).

Now consider dividing the regressors in the model into two subsets: one subset $\mathbf{x}_1$, whose coefficients are jointly of interest, and the complementary subset $\mathbf{x}_2$, with the remaining regressors. The generalized variance inflation factor (or GVIF) compares the squared size of the joint confidence region for coefficients associated with $\mathbf{x}_1$ with the squared size of the joint confidence region for similarly dispersed regressors that are uncorrelated with $\mathbf{x}_2$:

$$\text{GVIF}_1 = \frac{\det \mathbf{R}_{11} \det \mathbf{R}_{22}}{\det \mathbf{R}},$$

where $\mathbf{R}_{11}$ is the correlation matrix among the regressors in $\mathbf{x}_1$; $\mathbf{R}_{22}$ is the correlation matrix among the regressors in $\mathbf{x}_2$; $\mathbf{R}$ is the correlation matrix for all of the x s; and $\det$ is the matrix determinant. When there is only one regressor in $\mathbf{x}_1$, the GVIF reduces to the usual VIF.

—John Fox

REFERENCES

- Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12, 591–612.
- Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87, 178–183.

VARIATION

The term *variation* refers to the dispersion or SPREAD on a VARIABLE. There is no variation if all observations have the same value on the variable. There is variation whenever the observations are not all alike. Variation can be seen as the degree of heterogeneity on the variable.

STATISTICS would be an unnecessary field if variation did not exist. The interesting variable is one that has variation, and statistics seeks to assess the sources of that variation. CENTRAL TENDENCY measures show what value is typical for the variable, but variation measures are needed to show how typical that value is.

For numeric data, variation is defined as the sum of squared deviations from the MEAN (Blalock, 1979, p. 338). If X_i is the i th observation's value on variable X , and if $\bar{X}$ is the mean (see $\bar{X}$), then variation = $\sum (X_i - \bar{X})^2$. The VARIANCE of a variable is defined as its variation divided by the number of observations, N , and the STANDARD DEVIATION is defined as the square root of the variance.

The variation on a variable can be due to either systematic forces (such as other variables that affect the variable of interest) or to ERROR (such as error in measuring the variable). Many statistical techniques, such as REGRESSION and ANALYSIS OF VARIANCE, seek to partition the sources of variation in the DEPENDENT VARIABLE. In regression analysis, the total variation is divided into the variation due to regression and the variation due to error. If all the data points fall along

a straight line in LINEAR REGRESSION, for example, all of the variation is due to regression and none is due to error. The *R-SQUARED* value in regression is the ratio of the variation due to the regression to the total variation of the dependent variable. In analysis of variance, the total variation is partitioned into the portion due to each independent variable separately, possibly the portion due to INTERACTIONS of the independent variables, and unexplained variation. In using these techniques, researchers seek independent variables that minimize the amount of unexplained error variation, constrained by a desire for parsimony in terms of using the smallest possible efficient number of PREDICTOR VARIABLES.

For nonnumeric variables, variation involves examining the dispersion of values. For example, the VARIATION RATIO is one of several measures that can be used to assess the spread of observations on nominal data.

—Herbert F. Weisberg

REFERENCES

- Blalock, H. M. (1979). *Social statistics* (rev. 2nd ed.). New York: McGraw-Hill.
- Lewis-Beck, M. S. (1995). *Data analysis: An introduction*. Thousand Oaks, CA: Sage.

VARIATION (COEFFICIENT OF)

The *coefficient of variation* of a variable is measured as the ratio of the STANDARD DEVIATION of a *continuous* variable to the sample mean of the same variable. Paul D. Allison (1978) suggests that coefficient of variation can be used to compare *relative* variation of two or more groups. This coefficient is usually denoted by *CV* and can be defined as

$$CV = \frac{s_x}{\bar{X}}$$

In addition to conventional measures of sample variation, such as sample variance and standard deviation, researchers can also use the coefficients of variation of multiple groups to compare their *relative variability and homogeneity* with respect to a certain variable, say, *X*. A higher value in *CV* indicates that the group is relatively more heterogeneous. In contrast, a lower value in *CV* implies that the group is more homogeneous.

The rationale for using the coefficient of variation is that a conventional measure of variation, such as VARIANCE (or, equivalently, the standard deviation), is scale-dependent. Basically, if the data are rescaled, the variance and standard deviation are also rescaled. For example, if we double the value of a variable, thus doubling its variance, then the standard deviation also doubles. If we change the scale of annual income from dollars (e.g., 36,000) to thousands of dollars (e.g., 36.0), then the standard deviation will also decrease by a factor of 1000 (such as from 5,500 to 5.5).

The characteristic of scale dependence of standard deviation and variance makes them somewhat misleading for cross-group comparisons. The coefficient of variation provides a possible solution to this problem. Because both the numerator (i.e., the standard deviation of *X*, or s_x) and the denominator (i.e., the mean of *X*, or $\bar{X}$) of the definition of *CV* are scale sensitive, the scaling factor is cancelled out when calculating the coefficient of variation. As a consequence, *CV* is scale insensitive and usually reported as a ratio or a simple decimal number.

To illustrate the advantage of using the *CV* for cross-group comparisons, suppose a researcher is concerned about demographic homogeneity of two communities with respect to annual income. In one community, the mean income is \$26,000 with a standard deviation of \$3,000. In the other community, the average income is \$39,000 with a standard deviation of \$5,000. We find that a larger mean income comes with a larger standard deviation. Therefore, it is useful to assess the size of standard deviation of annual income *relative* to mean income when one attempts to compare these two groups. This leads to two coefficients of variation: $3000/26000 = 0.115$ for the first community and $5000/39000 = 0.128$ for the second community. Note that the difference between these two *CVs* is much smaller than that between the two standard deviations.

Although the coefficient of variation provides a measure for relative dispersion, this coefficient is reported rather infrequently in the social science literature. Jesper B. Sørensen (2002) points out some drawbacks associated with using the coefficient of variation. He argues that combining the mean and standard deviation into one measure sometimes may obscure the true effects of these two factors if both standard deviation and the mean have independent effects on the same response variable. He suggests researchers

use caution when using the coefficient of variation. Allison (1978) also provides more information regarding comparisons between several alternative measures of variation, such as the coefficient of variation and Theil's index.

—Cheng-Lung Wang

See also RELATIVE VARIATION (MEASURE OF)

REFERENCES

- Allison, P. D. (1978). Measures of inequality. *American Sociological Review*, 43, 865–880.
- Sørensen, J. B. (2002). The use and misuse of the coefficient of variation in organizational demography research. *Sociological Methods & Research*, 30, 475–491.

VARIATION RATIO

The variation ratio is a measure of DISPERSION for NOMINAL VARIABLES. It shows how descriptive the MODE is of the data. It is calculated as the proportion of cases that are *not* in its modal category. The variation ratio ranges from 0 to 1, with larger values indicating greater dispersion on the variable:

$$\text{variation ratio} = 1 - (f_{\text{mode}}/N),$$

where f_{mode} is the frequency of the modal category and N is the total number of cases.

The lower bound of the variation ratio is 0, attained when all cases are in the same category. Thus, zero values show that there is no dispersion on the variable.

The upper bound of the variation ratio is maximal when the mode is 1, meaning that each category has a frequency of 1, so there is complete dispersion on the variable. The variation ratio is then $1 - (1/N)$, which approaches 1 as the number of cases (each unique) is increased. For example, the variation ratio of possible nine-digit Social Security numbers is $1 - (1/1,000,000,000) = .999999999$, because each possible Social Security number is unique.

Say that a variable with k categories has a uniform DISTRIBUTION, with each category occurring with equal frequency N/k . The variation ratio is then equal to $1 - (1/k)$. Thus, its value under a uniform distribution is dependent on the variable's number of categories.

The advantage of the variation ratio as a measure of dispersion is that it is simple to compute. Its disadvantage is that it ignores much of the information in the data because it does not take the full distribution of cases into account. The index of diversity, index of qualitative variation, and information theory-based measures of entropy are measures of dispersion that are affected by the distribution of cases across all the categories (Weisberg, 1992, pp. 69–74).

As an example, consider the distribution of religious affiliation in various countries. About 60% of people in the United States are Protestants, so the variation ratio in the United States on religion would be 0.40. By contrast, the variation ratio would be only 0.10 in Poland because that country is about 90% Roman Catholic. Thus, the variation ratio shows that there is much greater dispersion on religion in the United States than in Poland. However, these values do not reflect the degree of dispersion among the other religions in these countries.

—Herbert F. Weisberg

REFERENCES

- Blalock, H. M. (1979). *Social statistics* (rev. 2nd ed.). New York: McGraw-Hill.
- Weisberg, H. F. (1992). *Central tendency and variability*. Newbury Park, CA: Sage.

VARIMAX ROTATION. See ROTATIONS

VECTOR

A vector is an entity in a geometric space that has both magnitude and direction. Graphically, the magnitude of the vector is commonly represented by the length of the line segment and the direction by an arrowhead at the end of the segment. Textually, vectors may be represented by their starting and ending points, often with an arrow above indicating direction. For example, $\overrightarrow{AB}$ indicates that the vector starts at point A and ends at point B .

In MATRIX ALGEBRA, a matrix having only one column is called a column vector. Similarly, a matrix having only one row is called a row vector. A common

use of vectors is in REGRESSION analysis (see ORDINARY LEAST SQUARES), where they are used to represent, for example, the set of n observations of the dependent variable in a $1 \times n$ column vector. EIGENVECTORS and their corresponding eigenvalues may also be used in REGRESSION DIAGNOSTICS (see also FACTOR ANALYSIS).

—Timothy M. Hagle

VECTOR AUTOREGRESSION (VAR).

See GRANGER CAUSALITY

VENN DIAGRAM

In a Venn diagram, overlapping areas of geometric shapes have meanings, facilitating comprehension of related concepts. Ruskey (2001) surveys the history of Venn diagrams and their technical details as a mathematical concept. Kennedy (2002) exposit their use in enhancing understanding of MULTIPLE REGRESSION ANALYSIS, the context in which they are relevant for social science research methods.

A classic Venn diagram is shown in Figure 1, in which the circle marked y represents “variation” in y , the dependent variable, and the circles marked x and w represent “variation” in the explanatory variables x and w , respectively. When y is regressed on x , the blue plus red area represents “information” used by the ORDINARY LEAST SQUARES (OLS) formula to produce its estimate of the slope coefficient of x . If such information represents variation in y uniquely explained by variation in x , the resulting estimate is UNBIASED. More information, reflected by a larger area of overlap, produces an estimate with a smaller variance (regardless of the “truth” of this information!). In the multiple regression context, when y is regressed on both x and w , the OLS estimator uses the information in the blue area to estimate the x slope coefficient, and the information in the green area to estimate the w slope coefficient, throwing away the information in the red area. The information in the red area is not employed because it does not correspond to variation in y uniquely attributable to either x or w . The result is unbiased estimates of the slopes of x and w .

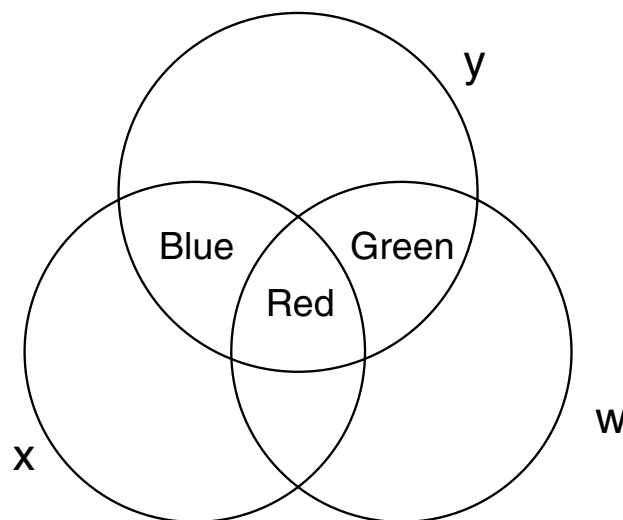


Figure 1 Venn Diagram in Which Colored Areas Represent Information Used by OLS Regression to Create Coefficient Estimates

This interpretation of the Venn diagram in a regression context can be used to exposit several results of interest, two examples of which follow.

1. Increasing the overlap between x and w increases the degree of MULTICOLLINEARITY. This causes the blue and green areas to shrink while the red area grows. Because the OLS formula continues to use the information in the blue and green areas to estimate the slopes of x and w , these estimates remain unbiased, but because there is now less information, their variances increase.

2. Omitting the relevant explanatory variable w will cause the OLS formula to use the blue plus red area to estimate the slope of x . This creates bias, due to inclusion of the red area, but because more information is used, variance decreases. An obvious exception is evident: No bias is created when x and w are orthogonal (i.e., no overlap, so no red area), such as might be the case in a designed EXPERIMENT.

—Peter E. Kennedy

REFERENCES

- Kennedy, P. E. (2002). More on Venn diagrams for regression. *Journal of Statistics Education*, 10(1) [online]. Retrieved from <http://www.amstat.org/publications/jse/v10n1/kennedy.html>

Ruskey, F. (2001). A survey of Venn diagrams. *The Electronic Journal of Combinatorics*, Dynamic Survey #5 [online]. Retrieved from <http://www.combinatorics.org/Surveys/ds5/VennEJC.html>

VERBAL PROTOCOL ANALYSIS

Verbal protocol analysis is typically a PILOT STUDY that seeks to do CONTENT ANALYSIS through systematic coding of verbalization of decision-making processes by the research subjects, especially consumers (Kuusela, Spence, & Kanto, 1998).

Verbal protocol analysis focuses directly on the sequence of cognitive events that occurs between the introduction of information stimuli and the decision outcome. Data are collected by setting up experimental manipulation or simply a task, introducing choice simulation, and then instructing subjects to report verbally all personal thoughts involved in the judgment/choice task. These thoughts are typically tape-recorded into protocols, then transcribed, broken into a sequence of task-relevant statements (protocol segments), and content analyzed using a CODING scheme with categories such as paired comparison of alternatives. The thoughts reflecting short-term memory processing and, thus, internal deliberation help to infer on issues such as choice criteria, decision strategies, and regularities. The protocol segments representing different aspects of the elementary information processes can also be analyzed qualitatively and quantitatively.

Verbal protocols, compared with other information-processing methods, provide more information per datum (Arch, Bettman, & Kakkar, 1978). Moreover, by tracing the decision-making process, they also explain it. This is useful in exploratory research, when there is no well-founded theory to guide the investigation of the target process. In addition, a simple coding of the protocol segments may serve as a basis for calculation of descriptive statistics relevant to theoretical issues, such as when and how often noncompensatory decision rules are used.

This methodology has some weaknesses. The need to provide a verbal report may alter what information is attended to in the stimulus (in particular, attention deficiency may arise in case of visual representations) and/or utilize some of the cognitive resources available

to the subject (mixing past and present, resulting in problems of accuracy and validity), thus fundamentally changing the underlying processes used in making a choice or judgment. There are also errors of transmission, commission, and omission. Transmission inaccuracy occurs because of an inadequate vocabulary or insufficient training in effective response capability. Commission error occurs where subjects systematically misreport their cognitive processes to impress the researcher by making their behavior look more rational. Omission occurs when the verbal reports are expected to be comprehensive, but the subject may distort or omit particular elements (prompted cues such as eye fixation patterns may help). The protocol method is also not capable of tracing cognitive processes that never reach consciousness. Such problems can be partially mitigated when protocol analysis is used to determine whether human behavior is governed by self-generated rules (Hayes, White, & Bissett, 1998). To tackle underlying cognitive processes one can also assess three criteria:

- *Relevance.* Participants should be talking about the task at hand, and not about an unrelated issue.
- *Consistency.* Verbalizations, to be pertinent, should be logically consistent with the verbalizations that just preceded them.
- *Memory.* A subset of the information heeded during the task performance should be remembered.

The protocol elicitation methods differ in terms of the following:

- The situation or decision task in relation to which protocol data are collected
- The amount of direction and guidance provided by the interviewer
- The timing of the protocol collection (concurrent vs. retrospective)

Concurrent protocols yielding more protocol segments than retrospective reflect the way information is processed (what, why, how). The focus being on *process*, this method is good for pilot studies. Retrospective protocol data, in the ideal case, are collected immediately after the task is completed, while most of the information is still in the short-term memory

and can be directly retrieved. Research shows that the subject here typically focuses on the *final* choice, and hence, this method is good for studies that focus on the *outcome* of the task.

—Pallab Paul and Kausiki Mukhopadhyay

REFERENCES

- Arch, D. C., Bettman, J. R., & Kakkar, P. (1978). Subjects' information processing in information display board studies. In K. Hunt (Ed.), *Advances in consumer research*, Vol. 5 (pp. 555–559). Ann Arbor, MI: Association for Consumer Research.
- Hayes, S. C., White, D., & Bissett, R. T. (1998). Protocol analysis and the “silent dog” method of analyzing the impact of self-generated rules. *Analysis of Verbal Behavior*, 15, 57–63.
- Kuusela, H., Spence, M. T., & Kanto, A. J. (1998). Expertise effects on prechoice decision processes and final outcomes: A protocol analysis. *European Journal of Marketing*, 32, 5–6, 559.

VERIFICATION

Verification refers to the use of empirical data, observation, test, or experiment to confirm the truth or rational justification of a hypothesis. Scientific beliefs must be evaluated and supported by empirical data. What does this require? Two concepts are fundamental in discussing scientific method: truth and justification (warrant). A hypothesis is true if it corresponds to the way the world is. Justification has to do with the grounds we have for believing a given statement to be true. A hypothesis is rationally warranted if a body of evidence and inference has been provided in support of it. Ideally, the fact that a statement is rationally warranted ought to make it likely that the statement is true. (This treatment makes apparent the relevance of Bayesian theory to scientific inference. The questions have to do with the transmission of rational credibility from one body of beliefs to another.) Philosophers have introduced several concepts and theories on the basis of which to analyze the logical and inferential relations between empirical data and scientific hypotheses, including observation, falsifiability, confirmation, and EXPERIMENTAL method. One of the most important consequences of this extended and complex debate is the conclusion that THEORIES cannot be “verified,” but they can be “confirmed,” “warranted,” or “falsified.”

The general question of scientific inference can be formulated in these terms: Given a body of evidence E and a HYPOTHESIS or theory T, how do we measure the warrant of T given E? Are there logical grounds for answering this question? (This does not define the full task for a theory of scientific method, because scientific method should also give us some guidance concerning the problem of discovering a relevant body of evidence.) It is rare for a scientific hypothesis to be amenable to direct and certain confirmation along these lines: Given E, H is certainly true. That is, it is rare that there is a finite body of observations that suffices to establish the truth of a given scientific hypothesis. This is so for two reasons: First, because scientific hypotheses normally refer to entities, mechanisms, or processes that are not directly observable; and second, because hypotheses and theories normally make universal claims (laws) that go beyond any finite body of observations. Instead, verification normally takes the form of indirect inductive or hypothetico-deductive support for the hypothesis: Given E, H is likely to be true.

Suppose that H deductively implies O, and that the body of evidence E contains “not O.” In this case, E falsifies H. On these assumptions, the body of evidence contains observations that demonstrate that H is false. (For example, suppose the hypothesis H deductively entails that the metal will melt when heated to 500 degrees centigrade. We perform the experiment, the metal does not melt, and we conclude that H is false.) Suppose now that H implies a series of observations $O_1, O_2, \dots, O_n$, and that E contains $\{O_1\}$. (That is, some of the observational consequences of the theory are found to be true.) Does E confirm H, and to what extent? This is the fundamental question of inductive logic and scientific method. What constitutes a significant test and confirmation to the theory? Several logical points are important. No finite list of observations exhaustively confirms a theory with universal generalizations, and different types of additional evidence have very different incremental effects on the credibility of the theory. For example, additional observations of the same type have less epistemic weight than an additional, unexpected observation.

What is an observation? This has been a main source of controversy in the philosophy of science, in that it has long been recognized that there is no sharp and permanent distinction between observation and theory. Virtually all scientific observations are theory-laden. But the essential idea is that an empirical observation

is a scientific belief with a relatively direct relationship between the evidence of the senses and the truth conditions of the statement, based on reliable techniques of data collection to which we can attach high rational credibility. Here, for example, we are to consider direct sensory observation, observation using instrumentation, interviews, records of price data, and so on. Most are not “direct” or “certain.” But they are relatively unburdened by currently controversial theories. This description suggests an approach to empirical confirmation that can be described as the “bootstrapping” method, where the credibility of some beliefs is enhanced by assigning provisional credibility to others; we then return to reassess the provisional beliefs based on the wider set of theoretical beliefs.

Several general approaches to empirical evaluation of scientific hypotheses have been offered in the past century. First is the idea, most deeply explored by Carl Hempel, that we should draw out the deductive consequences of a theory; evaluate the truth of some of those consequences using observation and instrumentation; and assign a degree of warrant to the theory based on the volume of its confirmed observational consequences. This approach constitutes the hypothetico-deductive theory of confirmation, and it represents the logical basis for the experimental method. Second is the idea, advanced by Karl Popper, that the scientific method works primarily through diligent and serious efforts to refute scientific hypotheses. The researcher needs to identify the most unlikely predictive consequences of the hypothesis and make every effort to show that these predictions are not upheld—thereby “falsifying” the theory in question. Only when a theory or hypothesis has withstood serious and varied tests along these lines do we have a rational basis for believing it. The two approaches are logically similar, in that they attempt to assign a degree of empirical warrant to a hypothesis based on the truth or falsity of its deductive consequences. But the underlying assumptions about how confirmation and testing proceed are quite different. Confirmation theory largely assumes that degree of warrant increases through the gradual accumulation of true consequences, whereas falsifiability theory assumes that warrant increases only as a result of our failure to refute the hypothesis in question. Both approaches have generated a large volume of critical discussion. Critical reflection upon classical confirmation theory notes that it is difficult or impossible to provide a purely formal specification of the set of deductive consequences of a theory that serves to

enhance the warrant of the theory. Further discussions of the theory of falsifiability have noted that anomalies are too easy to find in the development of scientific theories, and have attempted to offer a more historically adequate account of scientific belief based on a theory of “scientific research programmes” (Lakatos, 1978).

—Daniel Little

REFERENCES

- Brown, H. I. (1987). *Observation and objectivity*. New York: Oxford University Press.
- Glymour, C. N. (1980). *Theory and evidence*. Princeton, NJ: Princeton University Press.
- Hempel, C. (Ed.). (1965a). *Aspects of scientific explanation, and other essays in the philosophy of science*. New York: Free Press.
- Hempel, C. (1965b). Confirmation, induction, and rational belief. In C. Hempel (Ed.), *Aspects of scientific explanation*. New York: Free Press.
- Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers. Vol. 1*. Cambridge, UK: Cambridge University Press.
- Laudan, L. (1977). *Progress and its problems: Toward a theory of scientific growth*. Berkeley: University of California Press.
- Popper, K. (1965). *Conjectures and refutations: The growth of scientific knowledge* (2nd ed.). New York: Basic Books.

VERSTEHEN

This German word is usually translated as “to understand” and is often contrasted with “to explain” (*Erklären*). The contrast was used in 19th-century German discussions of the distinctive character of the *GEISTESWISSENSCHAFTEN*—the human sciences—by contrast with natural science. In this context, *verstehen* often refers to understanding from inside, rather than through study of external behavior, and it is often seen as the distinctive method of the human sciences. Equally important, what is produced is a different kind of knowledge from that generated by the natural sciences.

The original source of this idea was probably Vico’s suggestion that we can have deeper understanding of other human beings, including those who lived in the past, than we can ever have of physical phenomena. This is because we share a common human nature. However, this human nature is not a fixed set of

traits but an imaginative capacity, a capacity that both produced and can be used to understand past modes of thought and ways of life. Vico's work was largely neglected in the 18th century but was picked up by the German Romantics.

Dilthey is the most important 19th-century theorist of *verstehen*. He spent most of his long academic career seeking to clarify and develop this notion as the key method of the social and historical sciences. His earliest formulations portrayed it as a psychological process whereby historians project themselves imaginatively into the lives of people in the past, recreating their experience, so as to capture the meanings embodied in texts and artifacts. He believed there was a need for a philosophical psychology, or HERMENEUTICS, that would ground this aspect of historiography, and he set out to provide it. Dilthey's later accounts of *verstehen* portray it as cultural interpretation, rather than as a subjective process. It involves locating what is to be understood in terms of the objective cultural meanings that make up its context. Part of Dilthey's concern here was to recognize historical diversity while resisting epistemological RELATIVISM.

Dilthey's ideas have had considerable influence within the social sciences, notably through Max Weber's *verstehende* sociology (see IDEAL TYPE). However, they have also been subjected to serious challenge. In the 20th century, POSITIVISTS argued that, at best, *verstehen* could generate only hypotheses that still need to be tested, because it does not involve any rigorous form of validation. From the other side, there came criticism from Gadamer, who put forward what he called a philosophical hermeneutics. Where Dilthey saw *verstehen* as a method that is essential to the scientific character of social research, Gadamer treated it instead as a fundamental aspect of human-being-in-the-world. From this point of view, it does not differ from lay forms of understanding, nor does it have the procedural, fixed character of a method. Subsequently, some versions of STRUCTURALISM and POSTSTRUCTURALISM have questioned the very possibility of understanding in either sense.

—Martyn Hammersley

REFERENCES

- O'Hear, A. (Ed.). (1996). *Verstehen and humane understanding*. Supplement to *Philosophy*, Royal Institute of Philosophy Supplement 41. Cambridge, UK: Cambridge University Press.
- Palmer, R. E. (1969). *Hermeneutics*. Evanston, IL: Northwestern University Press.
- Truzzi, M. (1974). *Verstehen: Subjective understanding in the social sciences*. Reading, MA: Addison-Wesley.

VIGNETTE TECHNIQUE

Vignettes consist of stimuli presented to research participants. Their purpose is to selectively portray aspects of reality to which participants are asked to respond. They take many forms, including written and spoken narratives, visual imagery, video, and sound.

Like any research method, vignettes address clearly defined research questions, and their form and application will be directed by the research questions posed, the topics under study, and the kinds of participant groups involved. The simplicity of scenarios can help to identify, clarify, and disentangle the complexities of real-world processes. They also provide standardized contexts that can be put toward understanding people's responses to particular situations.

To illustrate one application of the vignette approach, consider Hughes's (1998) QUALITATIVE RESEARCH exploring drug injectors' infection risk behavior. Previous research demonstrates drug injectors' risk behavior to be heavily influenced by situated contexts. Vignettes were used to capture, albeit partially, some of these situations. A short story-book vignette was created, narrating risk behavior scenarios confronting two hypothetical drug injectors. Drug injectors' responses to the vignette were set within an in-depth INTERVIEW GUIDE, demonstrating the utility of vignettes in multimethod research. Hughes (1998) observed that drug injectors related well to the vignette; some enjoyed the story and felt confident to identify possible outcomes for the vignette characters. Furthermore, participants drew from a range of perspectives in responding, including for the vignette characters, their peers, and themselves. These different perspectives helped to minimize the potential SENSITIVE TOPICS of the research and SOCIAL DESIRABILITY BIAS. In application, the vignette exposed some of the constituents of risk behavior contained within the vignette scenarios, including what people considered ought to happen compared with what they believed would be realistic responses in drug injectors' lives.

Vignettes can never completely capture the realities of people's lives, and, like any research method,

their application is contingent on research purposes. Methodological debate on vignettes focuses on the differences between what people might actually do in real-life situations and responses elicited from selective representations of real life featured in the “vignette world.” Critiques of the method, including that of Parkinson and Manstead (1993), argue that vignette data can be understood only within the context of people’s responses to particular scenarios and do not allow GENERALIZATION to understanding real life. Other commentators, such as Loman and Larkin (1976), suggest that some vignette forms, such as video and film, relate better to real life than others, such as written narratives. Notwithstanding, vignettes play a valuable role in augmenting insights into the social world and are well suited to multimethod research. They contribute toward understanding people’s perceptions, beliefs, attitudes, and behavior throughout the social sciences.

—Rhidian Hughes

REFERENCES

- Hughes, R. (1998). Considering the vignette technique and its application to a study of drug injecting and HIV risk and safer behaviour. *Sociology of Health and Illness*, 20, 381–400.
- Loman, L. A., & Larkin, W. E. (1976). Rejection of the mentally ill: An experiment in labeling. *Sociological Quarterly*, 17, 555–560.
- Parkinson, B., & Manstead, A. S. R. (1993). Making sense of emotion in stories and social life. *Cognition and Emotion*, 7, 295–323.

VIRTUAL ETHNOGRAPHY

The term *virtual* ETHNOGRAPHY denotes PARTICIPANT OBSERVATION of cultural activities mediated by information and communications technologies, usually the Internet. These online activities are viewed as cultural interchanges in their own right, and the ethnographer is able to participate, observe, and produce holistic descriptions of cultural forms and processes, just as in ethnography in face-to-face settings. The ethnographer might study formation of identities, status hierarchies, power relations, cultural values, and norms of behavior. The ethnography may describe how an online setting, such as the lesbian bar described by Correll (1995), is brought into being by participants as a distinctive cultural space. Virtual ethnography can involve a qualitatively based immersive approach, or it

can extend to more complex mapping and quantification of interaction patterns. The field site might be an asynchronous discussion forum, such as a newsgroup or bulletin board, or a synchronous environment, such as a chat room or multiuser environment (text-based or graphical). In either case, the researcher will keep logs of the messages or interactions together with field notes of his or her observations and reactions.

The ability to observe interactions in many online settings without the permission or knowledge of participants raises an ethical dilemma. Many researchers would argue that consideration should be made of the sensitivity of the topic under discussion and the reasonable expectations of participants on the uses to which their words might be put, before venturing on an observational study. Other problems in the application of virtual ethnography include lack of access to corroborative evidence on informants’ offline lives, and the restriction of analysis to those who actively participate, with the reactions of those who simply read without sending messages (lurkers) remaining invisible (Paccagnella, 1997). This is not an issue when the aim is to describe online culture in its own right, but it is problematic when the aim is to address broader cultural phenomena. An approach drawing on more overt REFLEXIVITY would stress how much ethnographers can learn from their own experiences of interacting with informants through the technology, and it would place more emphasis on understanding experiences of the technological mediation of presence.

Virtual ethnography is closely related to the online community perspective, which holds that the social formations found in, for example, Internet newsgroups can be seen as communities in their own right (Baym, 2000). The social formations in online space are held to be deeply meaningful to participants, despite the lack of face-to-face interaction and potential absence of a long-term commitment. Hine (2000) argues that the online community perspective encourages conservative interpretations of the ethnographic approach. An alternative conception of virtual ethnography offers the potential for spatially innovative methodological approaches, adapting ethnographic sensibilities to explore cultural connections across a variety of online and offline settings. Virtual ethnography thus becomes a more broadly applicable methodological approach, encompassing a wider range of topics than the cultures of the Internet alone.

—Christine M. Hine

REFERENCES

- Baym, N. (2000). *Tune in, log on: Soaps, fandom and online community*. Thousand Oaks, CA: Sage.
- Correll, S. (1995). The ethnography of an electronic bar: The lesbian cafe. *Journal of Contemporary Ethnography*, 24(3), 270–298.
- Hine, C. (2000). *Virtual ethnography*. London: Sage.
- Paccagnella, L. (1997). Getting the seats of your pants dirty: Strategies for ethnographic research on virtual communities. *Journal of Computer Mediated Communication*, 3(1). Retrieved from <http://www.ascusc.org/jcmc/vol3/issue1/paccagnella.html>

VISUAL RESEARCH

Visual research covers a range of methods and practices that involve the use of visual media and technologies at all stages of research. Since the early 1990s, the use of the visual in social science research has increased dramatically for two reasons. First, theoretical shifts within the social sciences have led researchers to recognize that visual materials are not necessarily more subjective than verbal or written data. Second, new visual media and technologies of increasingly high quality, lower cost, and greater ease of use are available. Consequently, a growing number of researchers use video, photography, drawing, and multimedia in their work.

Although the term *visual research* has become commonplace, visual research is rarely purely visual, but involves the combination of visual and verbal or written methods of doing research and representing this work to others. Therefore, visual methods should not be used in isolation from established methods, but are a means of recognizing and accounting for the importance of visual experience, knowledge, and communication in everyday life and human perception. Visual research may consist of the following:

1. Analysis of the content, process of production, or uses of existing visual images, involving qualitative or quantitative CONTENT ANALYSIS (Rose, 2001).
2. Production of visual images as part of a research project to visually document activities and events, or in collaboration with informants to produce images that represent particular ideas, understandings, or worldviews (Banks, 2001; Pink, 2001).
3. Use of images in interviewing to elicit responses from informants (see Banks, 2001; Pink, 2001).
4. Visual observation of events and activities (Emmison & Smith, 2000).

Often, individual researchers mix different visual methods in the same project with other methods, such as analysis of written texts, IN-DEPTH INTERVIEWING, or PARTICIPANT OBSERVATION.

Visual research is represented in a number of ways, depending on the medium used and the research objectives. It includes using photographs in written publications; photographic exhibitions or essays; ethnographic FILM AND VIDEO; and interactive hypermedia projects using multiple media and published online, on CD-ROM, or on DVD.

THE HISTORY OF VISUAL RESEARCH

Visual methods were first used in sociology and anthropology in the early 20th century. This mainly involved what has been called early “ethnographic photography” of poverty in Europe, colonial subjects overseas, and the production of films about other cultures made during anthropological and exploratory expeditions. Much of the scientific image production of this period has recently been heavily critiqued as exoticising and part of the oppressive colonial project. However, these images are valuable research documents. Later in the 20th century, two landmark projects signify further development in visual research. *Nanook of the North* (1921) has been heralded as the first ethnographic film. Representing the everyday challenges and joys of the lives of North American Eskimos, *Nanook* set the scene for many of the debates that have followed in ethnographic filmmaking throughout the 20th century. In 1942, Mead and Bateson’s photographic study *Balinese Character* was published. An attempt to document Balinese behavior visually, this work fitted with the aims of social scientists of its times—to produce objective visual data about human activity. The question of whether visual images could be objective became a key issue in 20th-century debates.

In the 1980s and 1990s, a changing theoretical atmosphere lent momentum to developments in visual research. The idea that OBJECTIVITY could ever be truly achieved was challenged across the social sciences, particularly among qualitative researchers. Instead, the

notion of REFLEXIVITY—researchers' self-awareness of their subjectivity and how this affects the type of data produced—became increasingly important in QUALITATIVE RESEARCH. In this climate, it became clear that visual data were no more likely to be too subjective to produce reliable materials than any other data, and researchers could make convincing arguments for using the visual in their research and representation of their work.

CONTEMPORARY USES OF VISUAL RESEARCH

Contemporary visual researchers largely use photography, video, drawing, and hypermedia in their work. There are often no hard-and-fast rules regarding which visual methods or media are most suitable. Methods often have to be designed specifically to fit with wider projects and to suit the particular groups of people with which the researcher is working. For example, Banks describes how he photographed as part of participant observation during a festival in India. His informants showed him what was important to them at this event by asking him to photograph key individuals and activities. Through this, he confirmed that his own understanding of the importance of these people and actions was correct (Banks, 2001, pp. 45–47). I experienced similar “informant-directed photography” when researching the Spanish bullfight—my informants asked me to take photographs, through which I learned about my own and their understandings of the bullfight (Pink, 2001). The photographs produced under such circumstances may also be used in photo-elicitation interviews, which allow researchers and informants to discuss the meanings and content of visual materials. In-depth interviews may also be the context in which images are produced. For example, in research about people's relationships with their homes, I videotaped interviews that consisted of tours of each informant's home, led by the informant and guided by my checklist of questions. This allowed me to collaboratively explore with my informants the visual aspects of their home and the meanings that different objects, images, and areas of the home held for them. In these examples, visual methods are not used as distinct from established methods; rather, they are integrated as aspects of participant observation or in-depth interviewing.

However, this integration is not entirely straightforward. Using visual methods raises new ethical

issues. Unlike research represented only in written form, the anonymity of informants who appear in visual materials cannot be guaranteed. Therefore, such work requires that agreements are carefully made and thought out with informants and other stakeholders in the research before work begins.

Visual research methods are becoming increasingly popular among researchers. They are necessarily employed and understood slightly differently across the social sciences because in each instance, they are informed by the theoretical and methodological priorities of the discipline concerned.

—Sarah Pink

REFERENCES

- Banks, M. (2001). *Visual methods in social research*. London: Sage.
- Emmison, M., & Smith, P. (2000). *Researching the visual*. London: Sage.
- Pink, S. (2001). *Doing visual ethnography: Images, media and representation in research*. London: Sage.
- Rose, G. (2001). *Visual methodologies*. London: Sage.

VOLUNTEER SUBJECTS

Volunteer subjects (individuals who volunteer to participate in a research investigation) became a focus of attention in the 1960s and early 1970s, stemming from findings that volunteers may respond differently from how nonvolunteers in the general population would respond. Ralph L. Rosnow and Robert Rosenthal (1970) suggested that SYSTEMATIC ERRORS resulting from volunteer bias, although not easily discerned, may exert a subtle influence on experimental and nonexperimental social science research. Through experimental studies, as well as an integrative review of extant literature, Rosenthal and Rosnow (1975) showed that volunteers differed from nonvolunteers in at least 17 characteristics (including education, social class, intelligence, sociability, and approval motivation), and that volunteer subjects are overly sensitive and accommodating to DEMAND CHARACTERISTICS.

According to Rosnow and Rosenthal (1997, pp. 89–91), these findings have at least three practical implications for the social scientist. First, using only volunteer subjects may serve to weaken the explanatory power of the scientific model under study, because the

volunteer effect is usually not addressed explicitly as a variable in the model. A second implication has to do with the GENERALIZATION of research results to a population that typically includes nonvolunteers. A third implication pertains to ethical considerations as set out in the Belmont Report, developed in 1974 by a national commission in the United States charged with the task of formulating guidelines that would protect the rights and welfare of human research participants in biomedical and behavioral research (National Commission, 1979). These guidelines specified that only volunteer subjects could be used in most research investigations. In effect, this development implied that the volunteer problem could not be eliminated in RESEARCH DESIGN, only controlled or corrected for. Even so, precise estimation of the extent of BIAS remains a daunting task, because population parameters are rarely known. The best that the scientist can do in most instances is to speculate on the direction of the bias and interpret results accordingly. Fortunately, our knowledge of reliable characteristics of volunteers, and the reliability of volunteering for various types of studies, listed by Rosnow and Rosenthal (1997, pp. 95–98), can be used in this speculation.

An issue related to the volunteer subject is that of the pseudovolunteer, a verbal volunteer who fails to show up for a research appointment. Early studies had shown that research conclusions, including those examining

differences between volunteers and nonvolunteers, depend significantly on whether pseudovolunteers are included with the volunteer or the nonvolunteer group. Through a meta-analysis of research reports, Rosenthal and Rosnow (1975) found that, on average, approximately a third of those volunteering did not keep their appointments; more recent studies have confirmed that the MEDIAN pseudovolunteering rate has not changed noticeably since 1975.

—Ram N. Aditya and Ralph L. Rosnow

REFERENCES

- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. Washington, DC: Government Printing Office.
- Rosenthal, R., & Rosnow, R. L. (1975). *The volunteer subject*. New York: Wiley.
- Rosnow, R. L., & Rosenthal, R. (1970). Volunteer effects in behavioral research. In K. H. Craik, B. Kleinmuntz, R. L. Rosnow, R. Rosenthal, J. A. Cheyne, & R. H. Walters (Eds.), *New directions in psychology* (pp. 213–277). New York: Holt, Rinehart and Winston.
- Rosnow, R. L., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. New York: W. H. Freeman.

W

WALD-WOLFOWITZ TEST

The Wald-Wolfowitz (1940) test is a quick and easy test of the NULL HYPOTHESIS that two mutually independent random samples have been drawn from POPULATIONS that are identical. The test is completely general and is not very powerful against any specific kind of difference between the populations.

In order to carry out the test, the two sets of sample data, say m X s and n Y s, are pooled together and ordered from smallest to largest in a single array of $m + n = N$ observations, keeping track of which sample produced each observation. Then, the observations in the pooled array are replaced by X s and Y s that represent the respective populations of origin. If the X s and Y s are well mixed in this pooled array, there is no reason to think that the populations are different in any way. If the X s and Y s exhibit a pattern, this would suggest that there is a difference. For example, if most or all of the X s precede most of the Y s, this is evidence suggesting that the Y population tends to have values of larger magnitude than the X population.

In this pooled arrangement of m X s and n Y s, a run of X s is defined as a sequence of X s that is preceded and followed by one or more Y s or no letter at all. A run of Y s is defined similarly. The Wald-Wolfowitz test statistic T is the total number of runs of either X s or Y s. A large number of runs results when the X s and Y s are well mixed, and therefore, a large value of T tends to support the null hypothesis of identical populations. A small number of runs suggests that there is a pattern, and hence, the populations are not the same. Therefore, the null hypothesis should be rejected for small values

of T . Tables of the exact null distribution of T and/or CRITICAL VALUES of T are available in some books on nonparametric statistics. For larger sample sizes (m and n at least 10), an approximate test can be based on a standard NORMAL DISTRIBUTION with the following test statistic that is a function of T and incorporates a continuity correction to improve the accuracy:

$$Z = \frac{T + 0.5 - 1 - (2mn/N)}{\sqrt{\frac{2mn(2mn-N)}{N^2(N-1)}}}.$$

As an example, suppose that a random sample of six government economists (G) and seven academic economists (A) were asked to predict the overall percentage change in the CPI over the previous year. The data are as follows:

G: 3.1, 4.8, 2.3, 5.6, 0.1, 2.9

A: 4.4, 5.8, 3.9, 8.7, 6.3, 10.5, 10.8

Do the two sets of predictions come from the same population?

We pool the data into a single array as follows, with the government data values italicized.

0.1, 2.3, 2.9, *3.1*, 3.9, 4.4,
4.8, 5.6, 5.8, 6.3, 8.7, 10.5, 10.8

In terms of G and A, this arrangement is GGGGAAGGAAAAA, with a total of four runs, giving $T = 4$. The left-tail p VALUE from the table in Gibbons (1997, p. 483) for $m = 6$, $n = 7$ is 0.043. Using the large sample approximation, we obtain $Z = -1.73$. The left-tail p value from a table

of the standard normal distribution is 0.0418. This provides a very close approximation to the exact value. We reject the null hypothesis at the .05 level and conclude that the distributions are not the same.

Note that actual measurements are not necessary to carry out the test, as long as the pooled array can be formed. This Wald-Wolfowitz two-sample test is included in the STATXACT computer software package.

—Jean D. Gibbons

REFERENCES

- Gibbons, J. D. (1997). *Nonparametric methods for quantitative analysis* (3rd ed.). Columbus, OH: American Sciences Press.
- Wald, A., & Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Annals of Mathematical Statistics*, 11, 147–162.

WEIGHTED LEAST SQUARES

Weighted least squares (WLS) is an extension of ORDINARY LEAST SQUARES (OLS) regression that weights each observation according to some criterion, such as differences in the accuracy of estimates or error variances. In regular OLS, SLOPE estimates are found by minimizing the sum of squares, $\sum E_i^2$, where E represents the residuals from the regression. That is, we minimize the differences in the square of the actual values from the predicted values from the regression equation. WLS extends this model by minimizing the sum of the weighted least squares $\sum w_i^2 E_i^2$, where w represents a weight applied to each observation. Estimates from a WLS regression are interpreted in exactly the same manner as estimates from OLS. WLS is also efficient and shares with OLS the standard procedures for statistical inference (e.g., CONFIDENCE INTERVALS and hypothesis tests).

WLS is most frequently used in the social sciences when HETEROSKEDASTICITY (nonconstant error variance) is present in a regression analysis. The two problems associated with heteroskedasticity—inefficient estimates and biased standard errors—can be limited using weighted least squares. WLS can be used when the functional form of the relationship is linear or nonlinear.

Before WLS can be implemented to correct for heteroskedasticity, however, it must be known that the error variance is systematically related to one of the independent variables. Because this is not known before fitting a regression model, weights are typically found by examining the residuals from a preliminary OLS model. A common pattern of heteroskedasticity is the increase in error variance as the value of an independent variable, X , increases. In such cases, the variance of the errors is proportional to X , and thus, $1/X$ is usually an effective weight for computing the weighted least squares estimates. A plot of the log of the Studentized residuals against the fitted values can help detect this phenomenon.

The disadvantage of using WLS to correct for heteroskedasticity is that we require knowledge of the relationship between the errors and a particular independent variable. Because this cannot always be determined, OLS with robust standard errors is often used as an alternative. Unlike WLS, the use of robust standard errors does not change the OLS regression coefficients, but rather only increases the size of the standard errors. In other words, the use of OLS and robust standard errors will improve statistical testing, but it will not necessarily provide reliable coefficient estimates. As a result, if the pattern of heteroskedasticity can be determined, WLS is a better choice than OLS with robust standard errors.

When a nonrandom pattern in error variance occurs in a TIME SERIES, a related method, GENERALIZED LEAST SQUARES, is used in place of weighted least squares. Iterated weighted least squares (also called iteratively reweighted least squares) can also be used to fit GENERALIZED LINEAR MODELS. Finally, weighted least squares is also used in LOCAL REGRESSION to find local estimates and in ROBUST regression to lessen the impact of influential cases on the regression estimates.

—Robert Andersen

REFERENCES

- Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
- Neter, J., Wasserman, W., & Kutner, M. H. (1989). *Applied linear regression models* (2nd ed.). Boston: Irwin.

WEIGHTING

Weighting is a method used to estimate a characteristic of a population when data are collected from only a sample. Weighting a sample typically consists of attaching a multiplicative factor called a weight or case weight to each sample observation. An estimate of the population characteristic is the sum of the sample observations where each observation is counted proportional to its weight. In this sense, weighting is a convenient method for implementing a statistical estimation procedure to make inferences from a sample to the full population. Common synonyms for the weight are the *raising*, *inflation*, or *expansion factor*. These terms are descriptive because the weight inflates the sample to the population.

WEIGHTING FOR PROBABILITY SAMPLES

Some method of weighting is used to analyze the observations for almost all samples. Even unweighted analysis of a sample is a weighting scheme with all the case weights equal to unity. Weighting procedures differ depending on the application. With PROBABILITY SAMPLING, weighting the observations produces unbiased, or nearly unbiased, estimates of the population characteristic, where UNBIASED means the average of the estimates computed for all possible samples equals the population characteristic. For example, to compute an unbiased estimate of the population total from a simple random sample of size n from a population of size N , the weight for each observation is N/n . More generally, the reciprocal of the probability of selecting the unit in a probability sample is often called the base weight.

The base weight may be adjusted to compensate for units that do not respond or units that were not covered in sampling. For example, in TELEPHONE SURVEYS, the base weights may be adjusted to account for nonresponding households and for households without phones. Another reason for adjusting the base weights is to implement a more sophisticated estimation technique, such as ratio and regression estimators that take advantage of auxiliary data to improve the precision of the estimates.

WEIGHTING FOR OBSERVATIONAL SAMPLES

The main goal of weighting in observational samples selected without using probability sampling is also to reduce bias, but in this case, SELECTION BIAS is

the primary concern. For example, in an observational study of voters, the weights or balancing factors might be developed so the weighted sample matches the age, sex, and race of the population of registered voters. The objective of the weighting is to reduce the bias in the estimates that might result from having a sample that does not match the distribution of the population.

Standardization is an analytic tool frequently used to analyze epidemiological studies. Standardization is also used to make it easier to compare samples that have different compositions of members of the population. Kish (1987, pp. 113–115) discusses standardization as a method of weighting in observational studies to reduce selection biases. He shows that standardization is a weighting technique that eliminates the effect of unequal subgroup sample sizes and the bias associated with these unequal sample sizes. For example, suppose an equal number of male and female patients are treated for a disease, but the sample taken to evaluate the treatment contains more females than males. Weights may be attached to the sampled patients so that the weighted number of males equals the weighted number of females.

HISTORICAL DEVELOPMENT

Before high-speed computers were available, weighting using case weights was time-consuming and error-prone. As a result, self-weighting samples—in which every sampled unit has the same probability of selection and the same case weight—were very popular. With the introduction of computers, machine-readable cards containing the data from the observation were reproduced to properly represent units selected with unequal selection probabilities. For example, if some observations were sampled with a probability of $1/10$ and others were sampled with a probability of $1/20$, then the cards corresponding to those sampled at $1/20$ were duplicated.

The base weight for sampled unit i , w_i , is the reciprocal of the probability of selection and is a number between 1 and N . Consequently, it is natural to think of sampled unit i as representing w_i units in the population. Although this may be useful for understanding the idea of sampling and weighting, it loses its appeal when more advanced methods of weighting are considered. For example, a regression estimator is the base weight multiplied by an adjustment factor, and the adjusted weight may be less than zero for some cases. A negative regression weight does not mean that the sampled observation represents a negative number of

units in the population. In this situation, it may be better to consider the weight simply as a tool for calculating an estimate of the population.

EXAMPLE

Suppose a STRATIFIED simple random sample of schools is to be selected. The list of all schools in the population is divided into five strata based on the number of enrolled children. Every school in stratum h ($h = 1, 2, \dots, 5$) has a probability of selection equal to n_h/N_h , where N_h is the total number of schools in stratum h and n_h is the number of sampled schools in stratum h . The base weight for school i in stratum h is $w_{hi} = N_h/n_h$, the reciprocal of the selection probability. If only r_h sampled schools respond, a simple adjustment of the base weight is n_h/r_h , the reciprocal of the stratum response rate. The product of the base weight and the nonresponse adjustment is the nonresponse adjusted weight $w'_{hi} = N_h/r_h$ for responding school i in stratum h .

If the total number of public and private schools in the population is known from a census of the schools (say X_1 is the number of public schools and X_2 is the number of private schools), then a poststratified or ratio estimator can be computed. Weighting for the poststratified estimator is very similar to weighting for standardization. The adjustment factor for all responding public schools is $X_1/\hat{X}_1$, where $\hat{X}_1 = \sum w'_{hi}\delta_{hi}$ and $\delta_{hi} = 1$ if school i in stratum h is public, and it is zero otherwise. The factor for private schools is $X_2/\hat{X}_2$, where $\hat{X}_2 = \sum w'_{hi}(1 - \delta_{hi})$. The weights for the poststratified estimator are then $w''_{hi} = w'_{hi}X_k/\hat{X}_k$, where $k = 1$ for public schools and $k = 2$ for private schools. Because both the nonresponse adjusted weights and the poststratified weights could be used to produce estimates of the population, this example points out that the weighting scheme is not unique.

—J. Michael Brick

REFERENCES

- Babbie, E. (1979). *The practice of social research*. Belmont, CA: Wadsworth.
- Kalton, G. (1983). *Introduction to sampling*. Beverly Hills, CA: Sage.
- Kish, L. (1987). *Statistical design for research*. New York: Wiley.
- Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

WELCH TEST

The t -TEST of equality of means in two independent samples is relatively robust to violations of assumptions of variance homogeneity (HOMOSKEDASTICITY) and distribution normality within each sample (Wilcoxon, quoted in Algina, Oshina, & Lin, 1994; Winer, 1962). However, the t -test can result in wrong decisions if sample sizes and variances of the two populations are unequal. If the smaller sample has a larger variance, the ordinary t -test favors H_1 ; if the larger sample has a larger variance, H_0 is favored. The Welch test (Welch, 1947) corrects these deviations of ordinary t -test. If sample sizes are equal, the test statistic of the ordinary t -test and the modified t -test of Welch have the same values, but DEGREES OF FREEDOM differentiate. Welch's approximation generally results in lower degrees of freedom.

The test statistic of the Welch test is denoted by

$$t_{\text{Welch}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}.$$

Degrees of freedom are calculated by

$$df_{\text{Welch}} = \frac{1}{[c_1^2/(n_1 - 1)] + [c_2^2/(n_2 - 1)]},$$

where $c_i^2 = (s_i^2/n_i)/[(s_1^2/n_1) + (s_2^2/n_2)]$ for $i = 1$ or 2.

The Welch test should be used if variances are unequal (heteroskedasticity). The assumption of homoskedasticity can be tested by running Levene's F TEST, for example.

Illustration: Two groups are compared (see Table 1). Group 1 consists of 10 cases and has a mean of 10.0 in the analyzed variable X_1 . Group 2 consists of 40 cases. The mean of variable X_1 has a value of 12.0.

The variances in the two groups differ significantly (see Table 2). The value of Levene's F test is 8.5983

Table 1 Group Statistics

	Group	N	Mean	Std. Deviation
X_1	1.00	10	10.0263	0.9438
	2.00	40	11.9805	2.9635

NOTE: These are the results of a simulations study in which 100 simulations were run. For the first group, a normal distribution with mean 10 and variance 1 was assumed, and for the second group, a normal distribution with mean 12 and variance 3.

Table 2 Independent Samples Test

		<i>Levene's Test for Equality of Variances</i>		<i>t-Test for Equality of Means</i>		
		<i>F</i>	<i>Sig.</i>	<i>t</i>	<i>df</i>	<i>Sig. (2-tailed)</i>
X1	Equal variances assumed	8.5983	.0124	−2.0697	48	0.933
	Equal variances not assumed			−3.5417	42.9571	0.231

NOTE: These are the results of a simulations study. The specification of the study is described in Table 1. The values reported are averages. The F -value is the average over 100 simulations, the significance level of the F -value is the average over 100 simulations (and not of the value 8.5983), and so on.

(average of 100 simulations). The computed error level of 0.0124 (average of 100 simulations) is lower than the chosen threshold value of $p < .05$ for the significance level. Thus, the Welch test should be used, the results of which are noted in the second row of Table 2. The test statistic is equal to -3.5417 , and the degrees of freedom are 42.9571. The test statistic is significant if a threshold value of $p < .05$ is chosen for the significance level. In contrast, the ordinary t -test presents no significant results. The reason for this difference is that the ordinary t -test favors H_0 if the larger sample has a larger variance. The example is based on a simulation, and hence, the exact test statistic is known. It has a value of -3.51 and is significant for the selected threshold value of $p < .05$. However, the differences between the two tests dissipate if sample sizes are increased, the variance of Group 2 is reduced, or the number of cases is allocated more equally to the two groups. If, for example, Group 1 consists of 20 cases and Group 2 of 80, both tests result in significant differences [ordinary t -test: $t = -2.9842$ ($p < 5\%$); Welch test: $t_{\text{Welch}} = -5.0580$ ($p < 0.1\%$)].

The Welch test is commonly implemented in statistical software packages such as SPSS or SAS. The test becomes very conservative, favoring H_0 , if the underlying distributions get longer and the sample sizes decrease (e.g., from 20 to 10 in each group) (Yuen, 1974). Like the ordinary t -test, the Welch test can result in biases in both directions (favoring H_0 or H_1) if the underlying distributions are (very) skewed (Algina et al., 1994), especially if SKEWNESS in different directions causes problems. Depending on the amount and kind of skewness, combined or total sample sizes of 40, 80, or, in very extreme situations, 200 cases are needed in order to get sufficient results (Algina et al., 1994).

Good textbooks often recommend to test whether observed distributions in each group deviate from

normal distribution (see KOLMOGOROV-SMIRNOV TEST) if sample sizes are small, and to use a nonparametric test (e.g., MANN-WHITNEY U or WILCOXON rank sum test) if the assumption of normality is violated. A wrong interpretation of such advice, sometimes found in textbooks and research methods, is that the total distribution should be normal. Furthermore, deviations from normality do not necessarily equate with poor performance of the t -test or Welch test. Yuen (1974), for example, notes excellent Welch test results in the case of underlying uniform distributions for small sample sizes (e.g., for $n = 10$ for each group).

Skewness, rather than deviations from normality, seems to be an important factor. Therefore, it is recommended that one should test for skewness if combined sample sizes are smaller than 200, run a t -test or Welch test and a nonparametric test, and compare the results of both tests if data are skewed unless more experiences are available. In smaller samples, outliers may cause heteroskedasticity and skewness. Hence, the data should be checked for outliers.

—Johann Bacher

See also DIFFERENCE OF MEANS

REFERENCES

- Algina, J., Oshina, T. C., & Lin, W.-Y. (1994). Type I error rates for Welch's test and James's second order test under non-normality and inequality of variance when there are two groups. *Journal of Educational and Behavioral Statistics*, 19, 275–291.
- Welch, B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika*, 34, 28–35.
- Winer, B. J. (1962). *Statistical principles in experimental designs*. New York: McGraw-Hill.
- Yuen, K. K. (1974). The two-sample trimmed t for unequal populations variances. *Biometrika*, 61, 165–170.

WHITE NOISE

In general, white noise is an unsystematic variable representing unmeasured properties. The phrase is most often used in the study of TIME SERIES variables, which exhibit AUTOCORRELATION (i.e., current values are correlated with earlier values). The most common is a first-order process, where values at time t are correlated with values at time $t - 1$. An example might be the annual U.S. unemployment rate, 1955–2002. It is easy to imagine that the unemployment rate in the current year is closely predicted by the unemployment rate from the previous year. Of course, some time series may show, or be transformed to show, no autocorrelation at all, at any LAG. Then, lagged values do not predict subsequent values, and the variable appears driven by a STOCHASTIC, or random, process referred to as white noise. If the variable under study is actually a RESIDUAL from a REGRESSION equation, it might be useful to test its status as a white noise process. Bartlett has proposed a white noise test. Supposing the regression residual was found to be white noise, then the need for the usual autocorrelation corrections to achieve more efficient estimates would disappear.

—Michael S. Lewis-Beck

WILCOXON TEST

The Wilcoxon (1945) test, also known as the Wilcoxon signed-rank test, is the nonparametric counterpart of the Student's T -TEST for the mean in the case of one sample, and the paired-sample Student's t -test for the mean difference in the case of paired samples. We discuss it here in the context of paired samples, because that is its most frequent use.

Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be a RANDOM SAMPLE of n pairs drawn from a bivariate distribution. We want to make an inference about the mean difference between the X s and Y s in the POPULATION. The first step is to take the differences between the X and Y observations in each pair, $D_i = X_i - Y_i$. Because the difference between the X population mean and the Y population mean is the same as the mean μ_D of the $D = X - Y$ difference population, the NULL HYPOTHESIS is $H_0: \mu_D = 0$.

To carry out the Wilcoxon test, the absolute values of the differences $D_1, D_2, \dots, D_n$ are arranged from smallest to largest, keeping track of the original sign of each difference. Then, these absolute values are assigned integer ranks from 1 to n . The test statistic T is the sum of the ranks of those differences that were originally positive. A large value of T suggests that the mean of the X population is larger than the mean of the Y population and, hence, calls for rejection of H_0 in favor of the positive alternative $H_1: \mu_D > 0$. Alternatively, a P VALUE can be calculated as the right-tail probability of a value as large as or larger than the observed value of T . Tables of the exact null distribution of T and/or CRITICAL VALUES of T are available for small sample sizes in Gibbons (1993, 1997). For larger sample sizes, an approximate test can be based on a standard NORMAL DISTRIBUTION with the following test statistic that is a function of T and incorporates a continuity correction to improve the accuracy:

$$Z = \frac{T - 0.5 - [n(n+1)/4]}{\sqrt{n(n+1)(2n+1)/24}}.$$

Any differences that are equal to zero are ignored and n is reduced accordingly. If any of the absolute values of the differences are tied, each of those tied at any given value is assigned the same midrank, defined as the average of the ranks those differences would have been assigned if they were not tied.

As an example, suppose that a random sample of 10 people who drive automobiles was selected to see whether alcohol has an adverse effect on reaction time. Each driver's reaction time was measured both before and after drinking a beverage containing 2 ounces of 80-proof alcohol. The data, measured in seconds, are shown in Table 1. If the alcohol has an adverse effect on reaction times, the sample differences, After – Before, will tend to be large and positive. These differences, their absolute values, and ranks are also shown in Table 1.

The value of the test statistic is $T = 48.5$ with $n = 10$. The exact right-tail p value from Gibbons (1993, p. 68) is 0.0165. The large sample approximation gives $Z = 2.08$ with a left-tail p value of 0.0188. In both cases, the conclusion is to reject the null hypothesis and conclude that alcohol does have an adverse effect on reaction time.

The Wilcoxon test is valid under the assumption that the population distribution of differences is symmetric and continuous. The normal approximation is reasonably accurate for sample sizes of at least 10, and it is

Table 1 Calculation of Wilcoxon Test Statistic

Subject	Before	After	Difference	Absolute Value	Rank	Positive Rank	Negative Rank
1	0.68	0.73	0.05	0.05	5.5	5.5	
2	0.64	0.63	−0.01	0.01	1.0		−1.0
3	0.68	0.66	−0.02	0.02	2.0		−2.0
4	0.82	0.92	0.10	0.10	9.5	9.5	
5	0.58	0.68	0.10	0.10	9.5	9.5	
6	0.80	0.87	0.07	0.07	8.0	8.0	
7	0.72	0.77	0.05	0.05	5.5	5.5	
8	0.66	0.70	0.04	0.04	3.5	3.5	
9	0.88	0.84	−0.04	0.04	3.5		−3.5
10	0.73	0.79	0.06	0.06	7.0	7.0	
						48.5	

improved by incorporating a correction for ties in the Z statistic. The test is almost as powerful as Student's *t*-test for normal distributions and can be more powerful for other distributions. Many computer packages of statistical tests include this test.

A different test that is frequently called the Wilcoxon test is the Wilcoxon rank sum test for the null hypothesis of equal means in two mutually independent random samples and is equivalent to the so-called MANN-WHITNEY U-TEST.

—Jean D. Gibbons

REFERENCES

- Gibbons, J. D. (1993). *Nonparametric statistics: An introduction*. Newbury Park, CA: Sage.
- Gibbons, J. D. (1997). *Nonparametric methods for quantitative analysis* (3rd ed.). Columbus, OH: American Sciences Press.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80–83.

WinMAX

WinMAX is a computer software package for qualitative data analysis. It is capable of grounded theory-oriented qualitative analysis of text data as well as quantitative type CONTENT ANALYSIS. For further information, see the website of the software: www.scolari.co.uk/frame.html or www.scolari.co.uk/winmax/winmax.htm.

—Tim Futing Liao

WITHIN-SAMPLES SUM OF SQUARES

An ANALYSIS OF VARIANCE is used to determine whether there are differences in the means of a variable for different groups. This question can also be stated as whether there is an effect of a NOMINAL VARIABLE on a QUANTITATIVE VARIABLE. If the means are different, this must be due to the effect of some variable.

The key to the understanding of analysis of variance lies in the fact that the observations on the dependent variable are different from each other. If the observations were all the same, then there could not be any effects of any variables. If the nominal, independent variable were the only variable that had any effect on the DEPENDENT VARIABLE, then the observations in each group would be identical. If gender were the only variable that affected income, then all women would have the same income and all men would have the same income, which would be different from the income of the women. This can be simplified by saying that all women would have an income equal to the mean income of women, and all the men would have an income equal to the mean income for men.

But most of the time, the observed values of the dependent variable are different from each other and different from the mean for the group. The fact that not all women have the same income implies that there is one or more variables, other than gender, that also have an effect on income. Thus, it is the VARIATION in the observations within a group that tells us that there are effects of other variables present in the data. Furthermore, it is the magnitude of the variations that tell us whether the effects are small or large.

Next, how can we measure the magnitude of the variations of the observations within a group? One possibility is to see how much each observation deviates from the mean in the group, square the deviation from the mean, and then add the squares. For example, if the observations in a particular group equal 7, 9, 10, and 14, then the mean equals 10. The deviations from the mean become -3 , -1 , 0 , and 4 ; the squares of these numbers are 9 , 1 , 0 , and 16 ; and the sum of these squares equals 26 . That is the magnitude of the variation of the observations within this particular group, as measured by a sum of squares.

Next, we find the magnitudes of the variations in each group and add all these sums of squares across the groups. This sum is known as the within-sample sum of squares. It is also called just the within-group sum of squares, or simply the within sum of squares, or the error sum of squares. Because the variations within the groups are due to all variables other than the nominal variable that defines the groups, the sum of squares is also better known as the residual sum of squares.

The reasons for using squares, and not some other method, are mainly historical and mathematical.

—Gudmund R. Iversen

REFERENCES

- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Iversen, G. R., & Norpoth, H. (1987). *Analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series 07-001, 2nd ed.). Beverly Hills, CA: Sage.
- Moore, D. S., & McCabe, G. P. (1998). *Introduction to the practice of statistics*. New York: W. H. Freeman.

WITHIN-SUBJECT DESIGN

Before common availability of statistical software, EXPERIMENTAL DESIGNS were developed with ease of computations in mind. Definitional formulas were hard to use, and efforts were made to simplify the necessary computations. But that made it difficult to see similarities between designs. Within-subject analysis is very similar to REPEATED-MEASURES DESIGNS. Both can now be handled in their simplest case as a two-way ANALYSIS OF VARIANCE with one observation per cell.

The within-subject design is used for taking measurements under different conditions, such as test scores, from several subjects. The question is whether there is any change in these measurements. Change implies that the variation in the scores for the subjects due to the different conditions is larger than what can be expected from randomness alone.

We assess the overall variation in the measurements for a single subject by finding the mean value for the subject, subtracting this mean from each measurement, squaring the differences, and adding the squares. This gives the variability in the scores *within* a particular subject, thus the name of the design. After finding the within SUM OF SQUARES for each subject, we add these sums to get the overall within-subject variation. In symbols, the within-subject sum of squares equals

$$\sum \sum (y_{ij} - \bar{y}_i)^2,$$

where y_{ij} is the j th measurement on the i th subject and from which we subtract the mean for the i th subject. The summation is across all conditions and all subjects. With k measurements on each subject, the within sum of squares for a subject has $k - 1$ DEGREES OF FREEDOM. With data on n individuals, the within-subject sum of squares above has $n(k - 1)$ degrees of freedom.

There are two reasons for the within-subject variation. One is attributed to the effect of the different conditions. The other is attributed to the residual variable.

The value of the residual variable for the j th measurement on the i th subject is found from the following expression:

$$\begin{aligned} (y_{ij} - \bar{y}) - (\bar{y}_i - \bar{y}) - (\bar{y}_j - \bar{y}) \\ = y_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}. \end{aligned}$$

From the deviation of the measurement from the overall mean, subtract the effect of the subject (subject mean minus overall mean) and the effect of the particular condition (condition mean minus overall mean). After taking out these two effects, whatever is left is the effect of the residual variable. (We recognize the resulting expression as the interaction effect in a two-way analysis of variance, used as the residual effect when there is only one observation per cell.) The residual sum of squares has $(n - 1)(k - 1)$ degrees of freedom.

The sum of squares for the different conditions can now be found as the difference between the within-subject sum of squares and the RESIDUAL SUM OF

SQUARES. The degrees of freedom for this difference equals $n(k - 1) - (n - 1)(k - 1) = k - 1$.

If the condition sum of squares is large, then there has been a change in the measurements from one condition to the next. We assess the magnitude of this sum of squares by comparing its mean square with the residual mean square, resulting in a value of the F variable and thereby a p value.

—Gudmund R. Iversen

REFERENCES

- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Sheskin, D. J. (1997). *Handbook of parametric and non-parametric statistical procedures*. Boca Raton, FL: CRC Press.
- Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

WRITING

Like all language, writing is a constitutive force that creates a particular view of reality and inscribes particular values. Styles of writing reflect historically shifting paradigms; social scientific writing, like all other writing, is a sociohistorical construction, neither immutable nor dispassionate. Writing in the social sciences is a site of contestation. How one writes, what one writes, and for whom one writes are theoretical, ethical, and methodological issues, sometimes referred to as “the crisis in representation” (Clifford & Marcus, 1986).

HISTORICAL DEVELOPMENT

Beginning in the 17th century, the Western world of writing was divided into two kinds: literary and scientific. Literature was associated with fiction, rhetoric, subjectivity, and an ambiguous (interpretive) voice, whereas science was associated with fact, truth, objectivity, and an unambiguous voice. During the 18th century, assaults against literary writing grew severe. Jeremy Bentham proposed that the ideal language would have no words, only symbols; David Hume described poets as professional liars; and Samuel Johnson sought, through his dictionary, to fix the meanings of words in perpetuity. Into this linguistic world, the Marquis de Condorcet introduced the term

social science, contending that if a precise language about moral and social issues were adopted, erroneous thought and action would be almost impossible. By the 19th century, literature and science stood as two separate domains. Literature was aligned with “art” and “culture,” containing the values of taste, aesthetics, ethics, humanity, and morality, whereas science was aligned with precision and objectivity. Because science’s status was greater than that of literature, some literary writers attempted to make literature a part of science. By the late 19th century, “realism” dominated both scientific and fiction writing, the latter spearheaded by Honore de Balzac and Emile Zola.

As the 20th century unfolded, the relationship between social scientific writing and literary writing grew more complex. The demarcations between “fact/fiction,” “subjective/objective,” and “true/imagined” blurred. The “New Journalists” made themselves the subjects of their writing; English departments positioned novels by minorities and postcolonials as “ethnographic fiction”; and social scientists, troubled by the “crisis in representation,” began publishing their research in literary formats—poetry, stories, and dramas (see CREATIVE ANALYTICAL PRACTICE ETHNOGRAPHY). In the 1970s, the crossovers spawned oxymoronic genres: “creative nonfiction,” “faction,” “ethnographic drama,” and “true novel,” to name a few. In the 1980s, John Van Maanen (1988) identified three kinds of ETHNOGRAPHIC TALES—realist, confessional, and impressionistic. By 1990, within social science, the schism between those who adhered to the scientific model of writing and those who chose to supplement that model with tools from the literary world widened. The schism shows no signs of narrowing.

Most of the social scientists who have expressly chosen literary tools are QUALITATIVE RESEARCHERS. Unlike quantitative researchers, who can express their findings in tables and charts, qualitative researchers must depend upon words. Because a goal of contemporary qualitative research is to bring to life for the reader the world studied in all its complexities, literary tools (e.g., scene setting, dialogue, multiple points of view, tone) and formats (essay, poetry, letter, dialogues, etc.) other than the research paper’s conventions and format have proved helpful.

Although most qualitative researchers purposefully choose literary devices, some continue to write within the general conventions of social science. Many of these view writing simply as a “mopping-up” activity,

“deskwork” following the real work or “fieldwork.” They are committed to the science model, and many have turned to computer programs for help in analyzing their data. But, increasingly, qualitative researchers are defining writing itself as a way of “knowing”—not just “telling”—a method of discovery, inquiry, and analysis.

WRITING AS METHOD OF INQUIRY

Writing as a method of inquiry departs from standard social scientific practices. It offers an additional—or alternative—research practice. The goal of writing as a method of discovery is to learn new things about one’s project, one’s self, and the relationship between the two. With this goal in mind, writers explore different writing formats, voices, points of view, and tones. In some ways similar to a free-ranging conversation with a trusted friend, a conversation without a preconceived notion of its outcome, this writing exploration creates the conditions for new insights—insights that would not occur if one followed the canonical format.

Writing as method of discovery further coheres with the core of POSTSTRUCTURALISM, namely, the *doubt* that any one theory or method, discourse or genre, tradition or novelty has a universal and general claim to authoritative knowledge. Rather, poststructuralism encourages claims to local, historical, and partial knowledge. That knowledge, moreover, shifts depending upon the voice, format, tone, and so on of one’s writing, along with shifts in one’s life contexts. Knowing about one’s research and knowing about oneself are inextricably intertwined, and unavoidably partial knowledges. Self and social science are “twin-born.”

Poststructuralism supports three writing strategies. First, the writing “I” that is always present, whether acknowledged or not, should be reflexively explored and inquired into by the writer. The writer asks, for example, “Why am I writing this text? What are my power interests?” Second, because one’s subjectivity—the writing “I”—is created in a context of competing discourses and local squabbles,

one’s subjectivity is neither fixed nor rigid, but shifts. Consequently, what one writes is not immune to the writer’s life situations. Writers are encouraged to contextualize their research in other frames, such as disciplinary, departmental, and familial relationships (see Richardson, 1997). Third, because knowledge is partial, the writer cannot say everything in one text, in one voice, in one format. Rather, the writer is encouraged to explore the material, shape it and cut it, from multiple points of view and in different formats. Examples of multiple takes on ethnographic materials are: *Composing Ethnography: Alternative Forms of Qualitative Writing*, edited by Carolyn Ellis and Arthur P. Bochner; *Troubling the Angels: Women Living With HIV/AIDS*, by Patti Lather and Christine Smithies; *Reading Auschwitz*, by Mary D. Langerwey; *Tales of the Field: On Writing Ethnography*, by John Van Maanen; *Travels with Ernest: Crossing the Literary-Ethnographic Divide* by Laurel Richardson and Ernest Lockridge; and *A Thrice Told Tale: Feminism, Postmodernism and Ethnographic Responsibility*, by Margery Wolf. In that book, Wolf tells about her research in Taiwan through a fictionalized story, her field notes, and a traditional social science article.

—Laurel Richardson

REFERENCES

- Clifford, J., & Marcus, G. E. (Eds.). (1986). *Writing culture as cultural critique*. Berkeley: University of California Press.
- Ellis, C., & Bochner, A. P. (Eds.). (1996). *Composing ethnography: Alternative forms of qualitative writing*. Walnut Creek, CA: Altamira.
- Richardson, L. (1997). *Fields of play: Constructing an academic life*. New Brunswick, NJ: Rutgers University Press.
- Richardson, L., & Lockridge, E. (in press). *Travels with Ernest: Crossing the literary-ethnographic divide*. Walnut Creek, CA: Altamira.
- Van Maanen, J. (1988). *Tales of the field: On writing ethnography*. Chicago: University of Chicago Press.
- Wolf, M. (1992). *A thrice told tale: Feminism, postmodernism and ethnographic responsibility*. Stanford, CA: Stanford University Press.

X

X VARIABLE

An INDEPENDENT VARIABLE in REGRESSION analysis is termed an X variable because it is usually depicted graphically along the horizontal axis (“ x -axis”) of a SCATTERPLOT. It is sometimes termed a right-side variable, in that it appears on the right-hand side of the regression equation. The X variable represents the presumed cause in a cause-and-effect relationship, whereas the Y variable denotes the effect.

Take a standard regression equation of the form $Y = a + bX$. The X on the right-hand side of the equation represents an independent variable. Consider an analysis of the effect of population density on crime rates in cities. The population density would be the X variable, which might be found to affect the crime rate, the DEPENDENT VARIABLE.

Similarly, in a DIFFERENCE OF MEANS test, the X variable is the DICHOTOMOUS VARIABLE for which the mean of the dependent variable is calculated. In looking to see whether men and women differ significantly in their vote turnout, for example, gender is the X variable. The turnout rates of men and women would be calculated, and then a DIFFERENCE-OF-MEANS test would be used to check whether that difference is statistically significant.

Generalizing to CROSS-TABULATIONS, percentages generally are taken within categories of the X variable, with comparisons made across categories of the X variable in order to gauge the effect of the independent variable on the dependent variable. In comparing the rates of church attendance of different religious denominations, for example, denomination would be

the X variable. The proportion of Catholics who go to church regularly might be compared to the proportion of Protestants and to the proportion of adherents of other religions who go to religious services regularly.

—Herbert F. Weisberg

$\bar{X}$

$\bar{X}$ is the notation for the arithmetic MEAN of a VARIABLE. If the variable is denoted by X , the i th observation’s value on variable X is denoted by X_i , and N is the total number of observations, then $\bar{X} = \sum X_i / N$. Figure 1 shows a density plot of spending on elementary and secondary education in the 50 American states, with the mean indicated on the graph.

The MEAN is most appropriate as a MEASURE OF CENTRAL TENDENCY when the data are NUMERIC, though the MEDIAN might be more appropriate when there are OUTLIERS on the variable. The mean also can be computed for DICHOTOMOUS VARIABLES (scored as either “1” or “0”), in which case the mean is simply the proportion of cases in the “1” category. The mean is also often computed for ORDINAL MEASURES, though that use is debatable. For example, Gravetter and Wallnau (2000) argued against calculating the mean and other traditional statistics with ordinal variables because the distance between values, and hence the mean, is not well defined. However, Borgatta and Bohrnstedt (1980) instead argued that there are latent CONTINUOUS VARIABLES underlying ordinal variables, so results obtained with integer scoring would be fairly

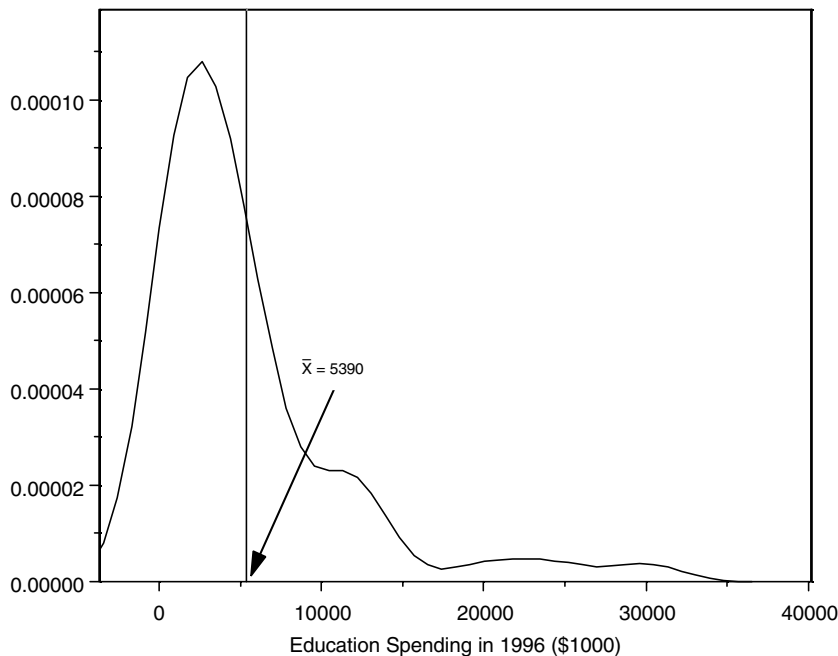


Figure 1 Density Plot With Mean Identified

SOURCE: State Politics & Policy Quarterly Data Resource at <http://www.unl.edu/sppq>

close to those that would be obtained for the true unknown numbered categories.

$\bar{X}$ has several distinctive properties leading to interpretations of the mean. First, the sum of signed DEVIATIONS around the mean is zero. That is $\sum (X_i - \bar{X}) = 0$. Furthermore, the sum of signed deviations around any value other than the mean is larger. As a result, the mean is considered a “best guess” statistic, in that guessing that the value of an observation is the mean leads to a smaller total error than would any other guess.

Second, the sum of positive deviations from the mean equals the sum of negative deviations from the mean (which is why the sum of signed deviations

around the mean is zero). Because of that property, the mean is sometimes described as the fulcrum (or balance point) of the variable’s distribution. To the extent that some values are above $\bar{X}$, they are balanced off by other values below $\bar{X}$.

Third, the sum of squared deviations around $\bar{X}$ is smaller than the sum of squared DEVIATIONS around any other value. This LEAST-SQUARES property of the mean is why VARIANCES are computed around the mean.

Additionally, $\bar{X}$ is a consistent, UNBIASED, and efficient estimator of the variable’s POPULATION mean. It is consistent because the probability that $\bar{X}$ minus the population mean is less than any value (δ) approaches 1 as the number of observations becomes infinitely large. It is unbiased because its expectation

is the population mean. $\bar{X}$ is considered efficient because the variance of SAMPLE means around their mean is smaller than it would be for any other estimator of the population mean.

—David C. Kimball and Herbert F. Weisberg

REFERENCES

- Borgatta, E. F., & Bohrnstedt, G. W. (1980). Level of measurement: Once over again. *Sociological Methods and Research*, 9, 147–160.
- Gravetter, F. J., & Wallnau, L. B. (2000). *Statistics for the behavioral sciences* (5th ed.). Belmont, CA: Wadsworth.

Y

Y-INTERCEPT

In a SIMPLE REGRESSION model, the y-intercept is the expected value of the DEPENDENT VARIABLE when the INDEPENDENT VARIABLE equals zero. In the regression equation $Y = a + bX$, a is the y-intercept. Visually, in a two-dimensional SCATTER PLOT with a regression line, the y-intercept is the value of y where the regression line crosses the vertical axis. In a MULTIPLE REGRESSION model, the intercept is the expected value of the dependent variable when all independent variables equal zero.

Consider the following example, in which a researcher is interested in factors that reduce heart disease, or at least reduce the most harmful effects of heart disease. In particular, this researcher wants to examine the RELATIONSHIP between wine consumption (X VARIABLE) and deaths from heart disease (Y VARIABLE). The main HYPOTHESIS is that increased wine consumption may lower the risks of heart disease. The researcher gathers data on annual wine consumption (liters per person) and heart disease deaths (per 100,000 population) from several countries. The estimated regression model from the data is $Y = 261 - 23X$. In this example, the y-intercept is 261, indicating that in a hypothetical country with no wine

consumption, the model predicts 261 deaths from heart disease for every 100,000 people in the country. The data are plotted in Figure 1 below, with the regression line included. The value of the y-intercept is noted in the figure.

Because $Y = a$ when $X = 0$, the y-intercept is also called the “constant” in the regression equation. In some substantive situations, the researcher may decide for theoretical reasons that the dependent variable should always be zero when the independent variable is zero. In that case, the regression equation can be calculated with no constant (so that a is zero), and the simple

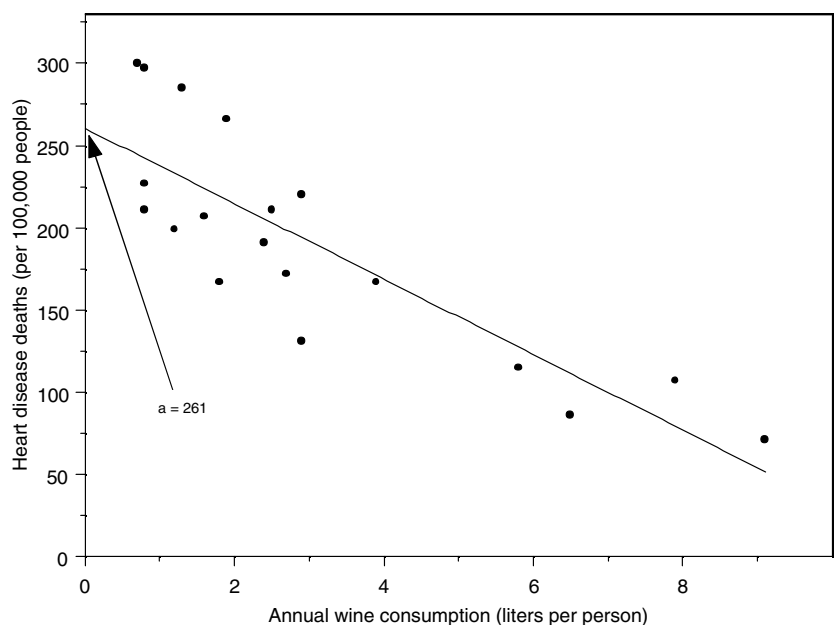


Figure 1 Scatterplot With y-Intercept Identified

SOURCE: Moore (1998), p. 314.

regression equation becomes $Y = bX$. Graphically, this would correspond to a regression line through the origin.

—David C. Kimball and
Herbert F. Weisberg

REFERENCES

- Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.
- Lewis-Beck, M. S. (1995). *Data analysis: An introduction*. Thousand Oaks, CA: Sage.
- Moore, D. S. (1998). *Statistics: Concepts and controversies* (4th ed.). New York: W. H. Freeman.

Y VARIABLE

The Y variable in regression analysis is the **DEPENDENT VARIABLE**, so named because it is conventionally depicted as the vertical (y) axis in a **SCATTERPLOT** showing the **RELATIONSHIP** between two variables. The Y variable represents the effect or outcome in a cause-and-effect relationship, whereas X usually denotes the cause or **INDEPENDENT VARIABLE**. The Y variable is generally shown on the left-hand side of regression equations; therefore, it is sometimes referred to as a left-side variable. For example, a **SIMPLE REGRESSION** equation is often written as $Y = a + bX + e$. Figure 1 shows a hypothetical relationship between years of education and income, where income is the Y variable because it is

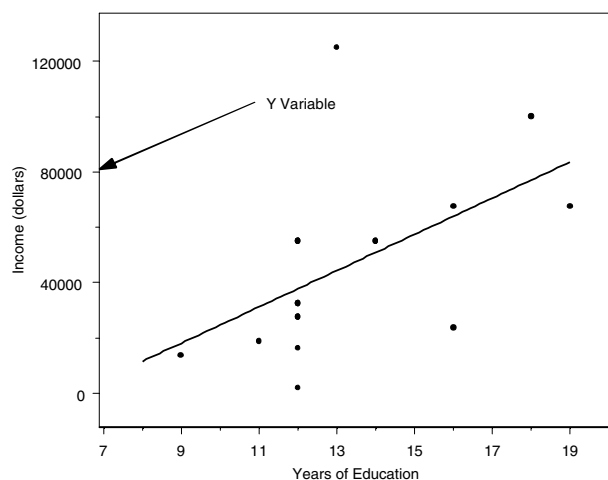


Figure 1 Scatterplot With Y Variable Identified

presumed to be affected by how many years of education a person has.

In **CROSS-TABULATIONS** of two variables, percentages are generally computed to determine the percentage of a given independent variable category that falls into each category on the dependent Y variable. In **DIFFERENCE-OF-MEANS** tests, the **MEAN** on the Y variable is computed within each category of the **PREDICTOR VARIABLE**, with a **SIGNIFICANCE TEST** to determine whether the obtained difference between the Y variable means is larger than expected on the basis of chance.

For example, in determining the relative contributions of different factors to a car's gas mileage, its miles per gallon would be the Y variable, whereas its weight, number of engine cylinders, and age might be plausible predictor variables. In this case, the **MULTIPLE REGRESSION** model could be written as $Y = a + bX + cW + dZ + e$, where Y represents the dependent variable (gas mileage) and X , W , and Z represent the predictor variables.

—David C. Kimball and Herbert F. Weisberg

YATES' CORRECTION

An adjustment used in connection with the **CHI-SQUARE** statistic. It is usually employed in connection with 2×2 **CONTINGENCY TABLES** and is sometimes used to improve the accuracy of tables with small **CELL** frequencies. It is used less frequently than in the past because of concerns that it may actually increase the chances of making a **TYPE II ERROR**. Essentially, the correction results in a smaller chi-square value and therefore a more conservative approach to chi-square, because it makes **STATISTICAL SIGNIFICANCE** more difficult to establish.

—Alan Bryman

YULE'S Q

A measure of **ASSOCIATION** used to indicate the strength of relationship between dichotomous variables at nominal or higher level. Yule's Q is also a

function of the ODDS RATIO standardized to fall between -1 and $+1$. To obtain Yule's Q from the odds ratio (OR) (Walsh & Ollenburger, 2001):

$$Q = (OR - 1)/(OR + 1).$$

Table 1 illustrates using Q with hypothetical social science data. In this example, the independent variable is gender and the dependent variable is political party. A working hypothesis from these data is that males are more likely to be Republicans than are females. Q is calculated from the cell frequencies in Table 1 using the odds ratio in the following manner:

$$OR = ad/bc,$$

$$OR = (10 \times 12)/(2 \times 4) = 120/8 = 15.$$

Q is then calculated as follows:

$$Q = (15 - 1)/(15 + 1) = 14/16 = .875.$$

Q is also calculated by finding the difference between concordant (P) and discordant (Q) pairs of data and dividing by the total of the concordant and discordant pairs. Tied pairs of data are excluded from the calculation. The number calculated from this method is the overall percentage of the concordant/discordant pairs and is also equal to the result calculated by the odds ratio method. The concordant/discordant formula for Q (Garson, 2002) is the following:

$$Q = (P - Q)/(P + Q).$$

Data pairs are concordant when moving from the first datum to the second involves a shift in the values of the independent variable and a shift in the values of the dependent variable consistent with the hypothesis being considered. Discordant pairs also involve shifts in value but ones inconsistent with the hypothesis. Tied data pairs share the same value of the dependent variable and are not part of the calculation.

$$Q = (ad - bc)/(ad + bc),$$

$$Q = [(10 \times 12) - (2 \times 4)]/[(10 \times 12) + (2 \times 4)],$$

$$Q = .875.$$

The results from these data indicate that there is a strong association between the variables in the data

Table 1 The Relationship of Gender to Choice of Political Party

	Male	Female
Republican	$a = 10$	$b = 2$
Democrat	$c = 4$	$d = 12$

set. Because the variables in this case are nominal, the sign of .875 is the result of the ordering of the variables. Reversing the order reverses the sign, but the magnitude of association remains the same. Q is a symmetric measure; it makes no difference which variable is chosen as the independent variable.

Q is interpreted as a proportionate reduction of error (PRE); for example, knowing someone's gender reduces our error in predicting the rank of the dependent variable by 87.5%. In other words, by knowing that a subject is male, one may guess the party affiliation is Republican and have a strong chance of being correct and supporting the working hypothesis. In the event Q equals 0, the relationship is statistically independent, so knowledge of the second variable is not helpful.

There is a note of caution in using Q . It is important to understand the distribution of the data and not rely totally on the measure itself. If any one of the frequency cells is 0, then calculating Q indicates there is a perfect relationship (White, 1999). Calculating Q from Table 2 in the same manner as before,

$$Q = [(10 \times 0) - (2 \times 4)]/[(10 \times 0) + (2 \times 4)],$$

$$Q = -1.0.$$

A perfect association of -1.0 is misleading in this particular case with only two female respondents in the data set.

Q is usually reported as GAMMA because there is a special application of it for 2×2 contingency tables. However, the value of Q is usually lower than gamma

Table 2 The Relationship of Gender to Choice of Political Party

	Male	Female
Republican	$a = 10$	$b = 2$
Democrat	$c = 4$	$d = 0$

because dichotomized data contribute to data loss and attenuation of the strength of association.

—Irvin B. Vann

Walsh, A., & Ollenburger, J. C. (2001). *Essential statistics for the social and behavioral sciences: A conceptual approach*. Upper Saddle River, NJ: Prentice Hall.

White, L. (1999). *Political analysis: Technique and practice* (4th ed.). Orlando, FL: Harcourt Brace.

REFERENCES

Garson, G. D. (2002, June). *Statnotes: An online textbook* [online]. Retrieved from <http://www2.chass.ncsu.edu/garson/pa765/statnote.htm>.

Z

Z-SCORE

A z -score is obtained by subtracting the MEAN from the value of a variable and then dividing the result by its STANDARD DEVIATION. Such a standardized score is the number of standard deviations that the value of a case is removed from the mean. It shows the relative status of that score in a DISTRIBUTION.

In social science research, we often replace original measurement scores by new ones, a procedure that is called TRANSFORMATION. The original scores are called *raw scores*, the new ones *transformed scores*. We may denote these as x and x' , respectively. If the transformation is linear, so that $x' = ax + b$, then it holds for the mean as well as the standard deviation that there will be a fixed relationship between the original value and the transformed value. For the mean, $\bar{x}' = a\bar{x} + b$, and for the standard deviation, $s_{x'} = as_x$. It follows that $a = s_{x'}/s_x$ and $b = \bar{x}' - a\bar{x} = \bar{x}' - (s_{x'}/s_x)\bar{x}$. Hence, the transformed scores can be written as $x' = ax + b = (s_{x'}/s_x)x + \bar{x}' - (s_{x'}/s_x)\bar{x} = (s_{x'}/s_x)(x - \bar{x}) + \bar{x}'$. In this formula, the mean and the standard deviation of the raw scores ($\bar{x}$ and s_x) are known from the data. Before transformation, we have to decide what the mean and standard deviation of the transformed scores will be. A popular choice is $\bar{x}' = 0$ and $s_{x'} = 1$. Then, the transformed score becomes $x' = (s_{x'}/s_x)(x - \bar{x}) + \bar{x}' = (1/s_x)(x - \bar{x}) + 0 = (x - \bar{x})/s_x$. This transformed score is known as the *standardized score* or *z -score*, and the transformation is called *z -transformation*: $z = (x - \bar{x})/s_x$. If we deal with the POPULATION instead of the sample, then Greek instead of Latin letters are used, so that $z = (x - \mu)/\sigma$.

AN EXAMPLE

Suppose you are studying the sleeping behavior of students, and you are told that François sleeps 10 hours per night. This score gives quite a different picture when the mean is 8 and the standard deviation is 2 than it does when the mean is 12 and the standard deviation is 4. In the first distribution, it is $z = (10 - 8)/2 = 1$, which means that François sleeps 1 standard deviation more than the mean, and in the second distribution, it is $z = (10 - 12)/4 = -0.5$, which means that he sleeps half a standard deviation less than the mean. So, he would be a sleepyhead in one distribution but a light sleeper in the other.

STANDARDIZATION OR NOT?

Standardization is useful because it leads to new sets of scores that are comparable. For example, the scores on two psychological tests will seldom be comparable. Changing to standardized scores permits comparison, so that it can be decided whether a person's performance on one test is better or worse than his performance on another. However, on the other hand, standardization is a procedure in which the variance of variables is artificially forced to be unity in order to obtain comparability, and it might be argued that this is too drastic an operation, because the score of an individual—for example, the educational level of François—has different meanings in societies with different amounts of educational variation. In this situation, nonstandardized scores (with indication of the standard deviation) might be preferred.

In general, standardized scores should be used when comparability is needed of many variables in one and

the same statistical model, and not used if the objective is to compare one or more variables of a statistical model in different samples. When we do not standardize, we must report the standard deviations of the variables.

OTHER WAYS OF ENSURING COMPARABILITY

A final remark is that standardization is but one way of expressing scores on different variables so as to have comparability. Another way would be the transformation to PERCENTILES.

It should also be noted that, in all the examples given above, variables were assumed to follow approximately a NORMAL DISTRIBUTION. For skew distributions and for comparison of variables with distributions of unequal form, the z -score becomes less valid as an indicator of the relative position of a variable in a distribution. In such cases, the nonnormal (e.g., skewed) distributions are transformed into a normal distribution. The transformation is then NORMALIZATION instead of linearization.

—Jacques Tacq

REFERENCES

- Blalock, H. M. (1972). *Social statistics*. Tokyo: McGraw-Hill Kogakusha.
- Hays, W. L. (1972). *Statistics for the social sciences*. New York: Holt International Edition.
- Tacq, J. J. A. (1997). *Multivariate analysis techniques in social science research: From problem to analysis*. London: Sage.
- Van den Ende, H. W. (1971). *Beschrijvende Statistiek voor Gedragwetenschappen*. Amsterdam/Brussel: Agon Elsevier.

Z-TEST

The z -test uses Z-SCORES to conduct SIGNIFICANCE TESTING for large samples. There are a number of significance tests in which the SAMPLING DISTRIBUTION has the form of the normal PROBABILITY DENSITY FUNCTION, which is then transformed into the standardized normal probability density function, so that the test scores become z -scores. A simple example is the test about a population mean such as a mean IQ score. If the NULL HYPOTHESIS states that the mean is 100 ($H_0 : \mu = 100$), then this is an exact hypothesis. If it states that the mean is at most 100 ($H_0 : \mu \leq 100$),

then it is an inexact hypothesis. The alternative hypothesis for the latter case could be that the mean is greater than 100 ($H_a : \mu > 100$). The test would then be one-tailed, because we hypothesize that the people of our sample are more intelligent than 100 on the average. For the two-tailed test, the alternative hypothesis would be $H_a : \mu \neq 100$. It is common to combine an inexact alternative hypothesis with an exact null hypothesis. For example, the alternative hypothesis that the mean IQ of a population is more than 100 ($H_a : \mu > 100$) is tested against the null hypothesis that the mean is equal to 100 ($H_0 : \mu = 100$).

Z-TEST OF A SINGLE MEAN

The z -test will be performed by evaluating the ratio $(\bar{x} - \mu) / \sigma_{\bar{x}}$ —where $\bar{x}$ is the sample MEAN, μ is the population mean under H_0 , and $\sigma_{\bar{x}}$ is the STANDARD ERROR of the mean (STANDARD DEVIATION of the sampling distribution of means under H_0)—when we are interested in testing the foregoing hypothesis about the mean IQ of the population. This ratio transforms a mean deviation into a z -score. Now suppose that in a sample of $n = 900$, randomly drawn with replacement, $\bar{x} = 102$ and $s = 10$. The sample is large enough, according to the CENTRAL LIMIT THEOREM, so the sampling distribution should be normal with mean equal to the population mean μ and variance $\sigma_{\bar{x}} = \sigma^2 / n$. Because σ is unknown, the variance of the sampling distribution will be estimated as $\hat{\sigma}_{\bar{x}}^2 = s^2 / n$ (instead of σ^2 / n), and its square root gives $\hat{\sigma}_{\bar{x}}$. Our sample result, $\bar{x} = 102$, is subsumed under the sampling distribution with mean 100 and variance $s^2 / n = 10^2 / 900 = 0.111$. Its z -score is then $z = (102 - 100) / \sqrt{0.111} = 6$. For this value of z or higher, we find in the standardized normal table a probability of 0.000001, which is so small (much smaller than a probability of making a TYPE I ERROR of $\alpha = .05$ or even of $\alpha = .001$) that it is highly unlikely to find this sample result under the null hypothesis. Hence, the null hypothesis is rejected, and we have found empirical support for the alternative hypothesis that $\mu > 100$. We may mention here that we do not know what the probability of TYPE II ERROR β is, but with a sample of large size ($n = 900$), we can be confident that β will be small.

OTHER APPLICATIONS OF THE Z-TEST

The z -test has many other applications. One is the test of a proportion. With presidential elections, the

proportion for determining winning is often $\pi = 0.50$. So, let us test the hypothesis that a candidate has more than half of the votes, that is, $H_a : \pi > 0.50$. This will be tested against the null hypothesis $H_0 : \pi = 0.50$. The proportion in the population is now indicated with the Greek letter π , and in the sample with the Latin letter p . The sampling distribution will now be a distribution of proportions (instead of means), given that the null hypothesis would be true. This sampling distribution has now the form of a BINOMIAL distribution. For large samples, it holds again that the NORMAL DISTRIBUTION can be used instead, because the binomial and normal probabilities then become identical. The mean of this sampling distribution of proportions is the population proportion π (under the null hypothesis), and its variance is $\pi(1 - \pi)/n$, where n is the sample size. Now suppose that in a sample of $n = 900$ before the elections, randomly drawn with replacement, we found $p = 53\%$ intended votes for the candidate. The sampling distribution of proportions, given a large sample, has the form of a normal distribution with mean $\pi = 0.50$ and variance $\pi(1 - \pi)/n = (0.50)(0.50)/900 = 0.0002777$. It then follows that $z = (p - \pi)/\sqrt{\pi(1 - \pi)/n} = (0.53 - 0.50)/\sqrt{0.0002777} = 1.8$. For this value of z or higher, we find in the standardized normal table a probability of .0359, which is smaller than $\alpha = .05$, so that it is highly unlikely we would find this sample result under the null hypothesis. Hence, with a likelihood of $1 - \alpha = 0.95$, the null hypothesis is rejected, and we have found empirical support for the alternative hypothesis that the candidate's votes should be significantly higher than 50%.

One remark of caution must be made. The binomial, a discrete distribution, is approximated here by the continuous normal distribution. Therefore, a value of p is, in fact, the midpoint of an interval, and a "correction for continuity" has to be made. In calculating the z -score, we subtract $0.5/n$ from p if p happens to be greater than π , and we add $0.5/n$ to p if p happens to be smaller than π . For our example, the z -score of the sample proportion p becomes $z = [p - (0.5/n) - \pi]/\sqrt{\pi(1 - \pi)/n} = [0.53 - (0.5/900) - 0.50]/\sqrt{0.0002777} = 1.767$, as opposed to the value of 1.8 obtained previously. The correction is rather small because of the large sample size. In general, the correction for continuity should be used especially when the sample size is small.

—Jacques Tacq

REFERENCES

- Blalock, H. M. (1972). *Social statistics*. Tokyo: McGraw-Hill Kogakusha.
- Hays, W. L. (1972). *Statistics for the social sciences*. New York: Holt International Edition.
- Tacq, J. J. A. (1997). *Multivariate analysis techniques in social science research: From problem to analysis*. London: Sage.
- Van den Ende, H. W. (1971). *Beschrijvende Statistiek voor Gedragwetenschappen*. Amsterdam/Brussel: Agon Elsevier.

ZERO-ORDER. See ORDER

Appendix

Bibliography

- Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, 16, 471–494.
- Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data. *American Journal of Sociology*, 96, 144–185.
- Abdi, H., Dowling, W. J., Valentin, D., Edelman, B., & Posamentier, M. (2002). *Experimental design and research methods*. Unpublished manuscript, University of Texas at Dallas, Program in Cognition and Neuroscience.
- Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks*. Thousand Oaks, CA: Sage.
- Abdi, H., Valentin, D., Edelman, B., & O'Toole, A. J. (1996). A Widrow-Hoff learning rule for a generalization of the linear auto-associator. *Journal of Mathematical Psychology*, 40, 175–182.
- Abell, P. (1987). *The syntax of social life: The theory and method of comparative narratives*. Oxford, UK: Clarendon.
- Abraham, B., & Ledolter, J. (1983). *Statistical methods for forecasting*. New York: Wiley.
- Abraham, B., & Ledolter, J. (1986). Forecast functions implied by ARIMA models and other related forecast procedures. *International Statistical Review*, 54, 51–66.
- Achen, C. H. (1975). Mass political attitudes and the survey response. *American Political Science Review*, 69, 1218–1231.
- Achen, C. H. (1977). Measuring representation: Perils of the correlation coefficient. *American Journal of Political Science*, 21, 805–815.
- Achen, C. H. (1982). *Interpreting and using regression*. Beverly Hills, CA: Sage.
- Achen, C., & Shively, W. P. (1995). *Cross-level inference*. Chicago: University of Chicago Press.
- Achinstein, P. (1983). *The nature of explanation*. New York: Oxford University Press.
- Achinstein, P., & Barker, S. F. (1969). *The legacy of logical positivism*. Baltimore: Johns Hopkins University Press.
- Adair, J. G., Sharpe, D., & Huynh, C. L. (1989). Hawthorne control procedures in educational experiments: A reconsideration of their use and effectiveness. *Review of Educational Research*, 59, 215–228.
- Aday, L. A. (1996). *Designing and conducting health surveys: A comprehensive guide* (2nd ed.). San Francisco: Jossey-Bass.
- Adler, P. A. (1985). *Wheeling and dealing*. New York: Columbia University Press.
- Adler, P. A., & Adler, P. (1991). *Backboards and blackboards*. New York: Columbia University Press.
- Adler, P. A., & Adler, P. (1994). Observational techniques. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 377–392). Thousand Oaks, CA: Sage.
- Adler, P. A., & Adler, P. (1998). *Peer power*. New Brunswick, NJ: Rutgers University Press.
- Adler, P., & Adler, P. A. (1987). *Membership roles in field research* (Qualitative Research Methods Series vol. 6). Newbury Park, CA: Sage.
- Adorno, T., & Horkheimer, M. (1979). *Dialectic of enlightenment* (J. Cumming, Trans.). London: Verso. (Original work published 1944).
- Afifi, A. A., & Clark, V. A. (1996). *Computer-aided multivariate analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall.
- Agar, M. (1973). *Ripping and running*. New York: Academic Press.
- Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Agresti, A. (1984). *An introduction to categorical data analysis*. New York: John Wiley.
- Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: John Wiley.
- Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley.
- Agresti, A. (1996). *An introduction to categorical data analysis* (2nd ed.). New York: John Wiley.
- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley Interscience.
- Agresti, A., & Cato, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician*, 54(4), 280–288.
- Agresti, A., & Finlay, B. (1997). *Statistical methods for social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Albrow, M. (1990). *Max Weber's construction of social theory*. London: Macmillan.
- Aldrich, J. H., & Cnudde, C. F. (1975). Probing the bounds of conventional wisdom: A comparison of regression, probit, and discriminant analysis. *American Journal of Political Science*, 19, 571–608.

- Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit and probit models*. Beverly Hills, CA: Sage.
- Algina, J., Oshima, T. C., & Lin, W.-Y. (1994). Type I error rates for Welch's test and James's second order test under non-normality and inequality of variance when there are two groups. *Journal of Educational and Behavioral Statistics*, 19, 275–291.
- Alkin, M. (1990). *Debates on evaluation*. Newbury Park, CA: Sage.
- Allison, P. D. (2003). Event history analysis. In M. A. Hardy & A. Bryman (Eds.), *Handbook of data analysis*. London: Sage.
- Allison, P. D. (1978). Measures of inequality. *American Sociological Review*, 43, 865–880.
- Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. C. Clogg (Ed.), *Sociological methodology 1987*. San Francisco: Jossey-Bass.
- Allison, P. D. (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Pine Forge.
- Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.
- Allport, G. (1942). *The use of personal documents in psychological science*. New York: Social Science Research Council.
- Alonso, E. (2002). AI and agents: State of the art. *AI Magazine*, 23(3), 25–29.
- Altheide, D. L. (1987). Ethnographic content analysis. *Qualitative Sociology*, 10, 65–77.
- Altheide, D. L. (1996). *Qualitative media analysis*. Newbury Park, CA: Sage.
- Altheide, D. L. (2002). *Creating fear: News and the construction of crisis*. Hawthorne, NY: Aldine de Gruyter.
- Altheide, D., & Johnson, J. (1994). Criteria for assessing interpretive validity in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 485–499). Thousand Oaks, CA: Sage.
- Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). *Statistics with confidence: Confidence intervals and statistical guidelines* (2nd ed.). London: British Medical Journal Books.
- Altman, M., Gill, J., & McDonald, M. P. (2003). *Numerical methods in statistical computing for the social sciences*. New York: Wiley.
- Alvarez, R. M., & Glasgow, G. (2000). Two-stage estimation of nonrecursive choice models. *Political Analysis*, 8, 147–165.
- Alveson, M., & Sköldbberg, K. (2000). *Reflexive methodology*. London: Sage.
- Alvesson, M. (2002). *Postmodernism and social research*. Buckingham, UK: Open University Press.
- Alvesson, M., & Deetz, S. (2000). *Doing critical management research*. London: Sage.
- Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MA: Harvard University Press.
- Amemiya, T. (1994). *Introduction to statistics and econometrics*. Cambridge, MA: Harvard University Press.
- American Association for Public Opinion Research (AAPOR). (2000). *Standard definitions: Final dispositions for case codes and outcome rates for surveys*. Ann Arbor, MI: Author.
- American Educational Research Association, American Psychological Association, & National Council on Educational Measurement. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist*, 47, 1597–1611.
- American Psychological Association. (2002, October 28). *PsycINFO database records* [electronic database]. Washington, DC: Author.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Anastasi, A., & Urbina, S. (1996). *Psychological testing* (7th ed.). New York: Prentice Hall.
- Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.
- Anderson, N. H. (1981). *Methods of information integration theory*. New York: Academic Press.
- Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandeale, W., & Weisberg, H. I. (1980). On the use of matrices in certain population mathematics. In *Statistical methods for comparative studies: Techniques for bias reduction*. New York: Wiley.
- Andersson, B. E., & Nilsson, S. G. (1964). Studies in the reliability and validity of the critical incident technique. *Journal of Applied Psychology*, 48(1), 398–403.
- Andrews, F. M., Morgan, J. N., & Sonquist, J. A. (1967). *Multiple classification analysis: A report on a computer program for multiple regression using categorical predictors*. Ann Arbor, MI: Survey Research Center, Institute for Social Research.
- Andrews, L., & Nelkin, D. (1998). Do the dead have interests? Policy issues of research after life. *American Journal of Law & Medicine*, 24, 261.
- Angle, J. (1986). The Surplus Theory of Social Stratification and the size distribution of personal wealth. *Social Forces*, 65, 293–326.
- Angle, J. (2002). The statistical signature of pervasive competition on wage and salary incomes. *Journal of Mathematical Sociology*, 26, 217–270.
- Angrasino, M. V., & Mays de Perez, K. (2000). Rethinking observation: From method to context. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 673–702). Thousand Oaks, CA: Sage.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444–455.

- Angst, D. B., & Deatrick, J. A. (1996). Involvement in health care decisions: Parents and children with chronic illness. *Journal of Family Nursing*, 2(2), 174–194.
- Anselin, L. (1988). *Spatial econometrics*. Boston: Kluwer Academic.
- Anselin, L., & Bera, A. (1998). Spatial dependence in linear regression models, with an introduction to spatial econometrics. In A. Ullah & D. Giles (Eds.), *Handbook of applied economic statistics* (pp. 237–289). New York: Marcel Dekker.
- Antoni, M. H., Baggett, L., Ironson, G., LaPerriere, A., August, S., Kilmas, N., et al. (1991). Cognitive-behavioral stress management intervention buffers distress responses and immunologic changes following notification of HIV-1 seropositivity. *Journal of Consulting and Clinical Psychology*, 59(6), 906–915.
- Arabie, P., & Boorman, S. A. (1973). Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology*, 10, 148–203.
- Arbuckle, J. (1997). *Amos users' guide version 3.6*. Chicago: Smallwaters Corporation.
- Arch, D. C., Bettman, J. R., & Kakkar, P. (1978). Subjects' information processing in information display board studies. In K. Hunt (Ed.), *Advances in consumer research*, Vol. 5 (pp. 555–559). Ann Arbor, MI: Association for Consumer Research.
- Archer, M. (1995). *Realist social theory: The morphogenetic approach*. Cambridge, UK: Cambridge University Press.
- Archer, M., Bhaskar, R., Collier, A., Lawson, T., & Norrie, A. (Eds.). (1998). *Critical realism: Essential readings*. London: Routledge Kegan Paul.
- Aristotle. (1977). *Metaphysica*. Baarn: Het Wereldvenster.
- Armstrong, R. L., Hosseini, J. C., Morris, S. A., & Rehbein, K. A. (1991). An empirical comparison of direct questioning, scenario, and randomized response methods for obtaining sensitive business information. *Decision Sciences*, 22, 1073–1090.
- Armitage, P., & Colton, T. (Eds.). (1998). *Encyclopedia of biostatistics*. New York: John Wiley.
- Armstrong, J. S. (Ed.). (2001). *Principles of forecasting*. Boston: Kluwer.
- Aronow, E., Reznikoff, M., & Moreland, K. (1994). *The Rorschach technique*. Needham Heights, MA: Allyn & Bacon.
- Aronson, E., Carlsmith, M., & Ellsworth, P. C. (1990). *Methods of research in social psychology*. New York: McGraw-Hill.
- Arrow, K. J. (1963). *Social choice and individual values* (2nd ed.). New Haven, CT: Yale University Press.
- Arthur, W. B. (1994). *Increasing returns and path dependence in the economy*. Ann Arbor: University of Michigan Press.
- Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 41, 258–290.
- Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership, and men* (pp. 177–190). Pittsburgh, PA: Carnegie Press.
- Asher, H. B. (1983). *Causal modeling* (2nd ed., Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–003). Beverly Hills, CA: Sage.
- Ashmore, M. (1989). *The reflexive thesis: Wrighting sociology of scientific knowledge*. Chicago: University of Chicago Press.
- Ashmore, M. (1995). Fraud by numbers: Quantitative rhetoric in the Piltdown forgery discovery. *South Atlantic Quarterly*, 94(2), 591–618.
- Atkinson, J. M., & Heritage, J. (Eds.). (1984). *Structures of social action: Studies in conversation analysis*. Cambridge, UK: Cambridge University Press.
- Atkinson, P. (1990). *The ethnographic imagination: Textual constructions of reality*. London: Routledge.
- Atkinson, P., & Hammersley, M. (1994). Ethnography and participant observation. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 248–261). Thousand Oaks, CA: Sage.
- Atkinson, P., & Silverman, D. (1997). Kundera's "immortality": The interview society and the invention of self. *Qualitative Inquiry*, 3(3), 304–325.
- Atkinson, P., Coffey, A., & Delamont, S. (2001). Editorial: A debate about our canon. *Qualitative Research*, 1, 5–21.
- Atkinson, R. (1998). *The life story interview* (Qualitative Research Methods Series, Vol. 44). Thousand Oaks, CA: Sage.
- Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.
- Axelrod, R. (1995). A model of the emergence of new political actors. In N. Gilbert & R. Conte (Eds.), *Artificial societies*. London: UCL.
- Axelrod, R. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton, NJ: Princeton University Press.
- Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211, 1390–1396.
- Ayer, A. J. (1936). *Language, truth and logic*. London: Victor Gollancz.
- Babbie, E. (1979). *The practice of social research*. Belmont, CA: Wadsworth.
- Babbie, E. R. (1995). *The practice of social research* (7th ed.). Belmont, CA: Wadsworth.
- Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, 73, 5–59.
- Bainbridge, E. E., Carley, K. M., Heise, D. R., Macy, M. W., Markovsky, B., & Skvoretz, J. (1994). Artificial social intelligence. *Annual Review of Sociology*, 20, 407–436.
- Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.
- Bakeman, R., & Quera, V. (1995). *Analyzing interaction: Sequential analysis with SDIS and GSEQ*. New York: Cambridge University Press.
- Baker, B. O., Hardyck, C. D., & Petrionovich, L. F. (1966). Weak measurement vs. strong statistics: An empirical critique of

- S.S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, 26, 291–309.
- Baltagi B. H. (2001). *Econometric analysis of panel data*. Chichester, UK: Wiley.
- Banerjee, A., Dolado, J. J., Galbraith, J., & Hendry, D. F. (1993). *Cointegration, error correction and the econometric analysis of nonstationary series*. New York: Oxford University Press.
- Banks, M. (2001). *Visual methods in social research*. London: Sage.
- Bargh, J. A., & Chartrand, T. L. (2000). Mind in the middle: A practical guide to priming and automaticity research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 253–285). New York: Cambridge University Press.
- Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948–955.
- Barndt, D. (1997). Zooming out/zooming in: Visualizing globalization. *Visual Sociology*, 12(2), 5–32.
- Barnes, D. B., Taylor-Brown, S., & Weiner, L. (1997). "I didn't leave y'all on purpose": HIV-infected mothers' videotaped legacies for their children. *Qualitative Sociology*, 20(1), 7–32.
- Barnett, V. (1974). *Elements of sampling theory*. Sevenoaks, UK: Hodder & Stoughton.
- Barr, R. (1996). A comparison of aspects of the US and UK censuses of population. *Transactions in GIS*, 1, 49–60.
- Barthes, R. (1967). *Elements of semiology*. London: Jonathan Cape.
- Barthes, R. (1977). *Image music text*. London: Fontana.
- Barthes, R. (1977). Rhetoric of the image. In *Image, music, text* (S. Heath, Trans.). London: Fontana.
- Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Arnold.
- Bartlett, M. S. (1947). Multivariate analysis. *Journal of the Royal Statistical Society, Series B*, 9, 176–197.
- Bartunek, J. M., & Louis, M. R. (1996). *Insider-outsider team research*. Thousand Oaks, CA: Sage.
- Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: John Wiley.
- Bateson, G., & Mead, M. (1942). *Balinese character: A photographic analysis*. New York: New York Graphic Society.
- Baudrillard, J. (1998). *The consumer society*. London: Sage. (Originally published in 1970)
- Bauman, Z. (1978). *Hermeneutics and social science*. London: Hutchinson.
- Baumrind, D. (1964). Some thoughts on ethics after reading Milgram's "Behavioral study of obedience." *American Psychologist*, 19, 421–423.
- Bayer, A. E., & Smart, J. C. (1991). Career publication patterns and collaborative "styles" in American academic science. *Journal of Higher Education*, 62, 613–636.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53, 370–418.
- Bayes, T. (1958). An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45, 293–315.
- Baym, N. (2000). *Tune in, log on: Soaps, fandom and online community*. Thousand Oaks, CA: Sage.
- Beck, N. (1991). Comparing dynamic specifications: The case of presidential approval. *Political Analysis*, 3, 51–87.
- Beck, N., & Jackman, S. (1998). Beyond linearity by default: Generalized additive models. *American Journal of Political Science*, 42, 596–627.
- Beck, N., & Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, 89, 634–647.
- Becker, H. S. (1967). Whose side are we on? *Social Problems*, 14, 239–247.
- Becker, H. S. (1974). Photography and sociology. *Studies in the Anthropology of Visual Communication*, 1(1), 3–26.
- Becker, H. S. (1986). *Doing things together*. Evanston, IL: Northwestern University Press.
- Becker, H. S. (1998). *Tricks of the trade*. Chicago: University of Chicago Press.
- Becker, H. S., Hughes, E. C., & Strauss, A. L. (1961). *Boys in white*. Chicago: University of Chicago Press.
- Becker, S., & Bryman, A. (Eds.). (2004). *Understanding research for social policy and practice: Themes, methods and approaches*. Bristol, UK: Policy Press.
- Bell, C. (1969). A note on participant observation. *Sociology*, 3, 417–418.
- Bell, S. E. (1999). Narratives and lives: Women's health politics and the diagnosis of cancer for DES daughters. *Narrative Inquiry*, 9(2), 1–43.
- Bellman, B. L., & Jules-Rosette, B. (1977). *A paradigm for looking: Cross-cultural research with visual media*. Norwood, NJ: Ablex.
- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley.
- Beltrami, E. J. (1999). *What is random? Chance and order in mathematics and life*. New York: Copernicus.
- Benfer, R. A., Brent, E. E., Jr., & Furbee, L. (1991). *Expert systems*. Newbury Park, CA: Sage.
- Bennet, D. J. (1998). *Randomness*. Cambridge, MA: Harvard University Press.
- Bennett, A. (1999). *Condemned to repetition? The rise, fall, and reprise of Soviet-Russian military interventionism, 1973–1996*. Cambridge: MIT Press.
- Benton, T. (1977). *Philosophical foundations of the three sociologies*. London: Routledge Kegan Paul.
- Benton, T. (1998). Realism and social science. In M. Archer, R. Bhaskar, A. Collier, T. Lawson, & A. Norrie (Eds.), *Critical realism: Essential readings*. London: Routledge Kegan Paul.
- Benzécri, J.-P. (1973). *Analyse des Données. Tome 2: Analyse des Correspondances* [Data analysis: Vol. 2. Correspondence analysis]. Paris: Dunod.

- Berelson, B. (1954). Content analysis. In G. Lindzey (Ed.), *Handbook of social psychology* (Vol. 1, pp. 488–522). Reading, MA: Addison-Wesley.
- Berg, B. L. (2001). *Qualitative research methods for the social sciences*. Boston: Allyn and Bacon.
- Berger, J. O., & Wolpert, R. L. (1984). *The likelihood principle: A review, generalizations, and statistical implications*. Hayward, CA: Institute of Mathematical Statistics.
- Berger, J., Fisek, H., Norman, R., & Zelditch, M. (1977). *Status characteristics and social interaction: An expectation states approach*. New York: Elsevier.
- Berger, P., & Luckmann, T. (1967). *The social construction of reality*. Harmondsworth, UK: Penguin.
- Bergner, R. M. (1974). The development and evaluation of a training videotape for the resolution of marital conflict. *Dissertation Abstracts International*, 34, 3485B. (UMI No. 73–32510)
- Bergsma, W. (1997). *Marginal models for categorical data*. Tilburg, The Netherlands: Tilburg University Press.
- Berk, R. A., & Freedman, D. A. (1995). Statistical assumptions as empirical commitments. In T. G. Blomberg & S. Cohen (Eds.), *Law, punishment, and social control: Essays in honor of Sheldon Messinger* (pp. 245–258). New York: Aldine de Gruyter.
- Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.
- Bernstein, R. (1983). *Beyond objectivism and relativism: Science, hermeneutics and praxis*. Oxford, UK: Basil Blackwell.
- Berry, J. (2002). Validity and reliability issues in elite interviewing. *PS—Political Science and Politics*, 35, 679–682.
- Berry, W. D. (1984). *Nonrecursive causal models*. Beverly Hills, CA: Sage.
- Berry, W. D. (1993). *Understanding regression assumptions*. Newbury Park, CA: Sage.
- Beverly, J. (1989). The margin at the center: On testimonio. *Modern Fiction Studies*, 35(1), 11–28.
- Beverly, J., & Zimmerman, M. (1990). *Literature and politics in the Central American revolutions*. Austin: University of Texas Press.
- Bhaskar, R. (1978). *A realist theory of science*. Hassocks, UK: Harvester Press.
- Bhaskar, R. (1979). *The possibilities of naturalism: A philosophical critique of the contemporary human sciences*. Brighton, UK: Harvester.
- Bhaskar, R. (1986). *Scientific realism and human emancipation*. London: Verso.
- Bhaskar, R. (1997). *A realist theory of science*. London: Verso. (Original work published 1978)
- Bhaskar, R. (1998). *The possibility of naturalism* (2nd ed.). Hemel Hempstead, UK: Harvester Wheatsheaf. (Original work published 1979)
- Bickman, L., & Rog, D. J. (Eds.). (1998). *Handbook of applied social research methods*. Thousand Oaks, CA: Sage.
- Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (Eds.). (1991). *Measurement errors in surveys*. New York: Wiley.
- Biggs, D., De Ville, B., & Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18, 49–62.
- Bijleveld, C. C. J. H., & van der Kamp, L. J. Th. (with Mooijaart, A., van der Kloot, W. A., van der Leeden, R., & van der Burg, E.). (1998). *Longitudinal data analysis: Designs, models, and methods*. London: Sage.
- Billig, M. (1996). *Arguing and thinking*. Cambridge, UK: Cambridge University Press.
- Bimler, D., & Kirkland, J. (1999). Capturing images in a net: Perceptual modeling of product descriptors using sorting data. *Marketing Bulletin*, 10, 11–23.
- Birdwhistell, R. L. (1970). *Kinesics and context: Essays on body motion communication*. Philadelphia: University of Pennsylvania Press.
- Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Birnbaum, M. H. (2001). *Introduction to behavioral research on the Internet*. Upper Saddle River, NJ: Prentice Hall.
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.
- Bishop, G., & Smith, A. (2001). Response-order effects and the early Gallup split-ballots. *Public Opinion Quarterly*, 65, 479–505.
- Black, D. (1958). *The theory of committees and elections*. Cambridge, UK: Cambridge University Press.
- Black, T. R. (1999). *Doing quantitative research in the social sciences: An integrated approach to research design, measurement, and statistics*. London: Sage.
- Blaikie, N. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity.
- Blaikie, N. (2000). *Designing social research: The logic of anticipation*. Cambridge, UK: Polity.
- Blalock, H. (1961). *Causal inference in nonexperimental research*. Chapel Hill: University of North Carolina Press.
- Blalock, H. M. (1960). *Social statistics*. New York: McGraw-Hill.
- Blalock, H. M. (1961). Theory, measurement and replication in the social sciences. *American Journal of Sociology*, 66(1), 342–347.
- Blalock, H. M. (1972). *Social statistics* (2nd ed.). Tokyo: McGraw-Hill Kogakusha.
- Blalock, H. M. (1979). *Social statistics* (rev. 2nd ed.). New York: McGraw-Hill.
- Blalock, H. M., Jr. (1969). *Theory construction: From verbal to mathematical formulations*. Englewood Cliffs, NJ: Prentice Hall.
- Blanck, P. D. (Ed.). (1993). *Interpersonal expectations: Theory, research, and applications*. New York: Cambridge University Press.

- Blascovich, J. (2000). Psychophysiological indexes of psychological processes. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 117–137). Cambridge, UK: Cambridge University Press.
- Blascovich, J., & Tomaka, J. (1996). The biopsychosocial model of arousal regulation. In M. Zanna (Ed.), *Advances in experimental social psychology* (pp. 1–51). New York: Academic Press.
- Blau, P. M. (1974). Parameters of social structure. *American Sociological Review*, 39, 615–635.
- Bloor, M. (1997). Addressing social problems through qualitative research. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (pp. 221–238). London: Sage.
- Blossfeld, H.-P., & Rohwer, G. (2002). *Techniques of event history modeling: New approaches to causal analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Blumen, I. M., Kogan, M., & McCarthy, P. J. (1955). *The industrial mobility of labor as a probability process*. Ithaca, NY: Cornell University Press.
- Blumen, I., Kogan, M., & McCarthy, P. J. (1968). Probability models for mobility. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in mathematical social science* (pp. 318–334). Cambridge: MIT Press.
- Blumer, H. (1956). Sociological analysis and the “variable.” *American Sociological Review*, 21, 683–690.
- Blumer, H. (1969). *An empirical appraisal of Thomas and Znaniecki (1918–20) The Polish Peasant in Europe and America*. New Brunswick, NJ: Transaction Books. (Original work published 1939)
- Blumer, H. (1969). *Symbolic interactionism: Perspective and method*. Englewood Cliffs, NJ: Prentice Hall.
- Blundell, R. W., & Smith, R. J. (1989). Estimation in a class of simultaneous equation limited dependent variable models. *Review of Economic Studies*, 56, 37–58.
- Blurton Jones, N. (Ed.). (1972). *Ethological studies of child behaviour*. London: Cambridge University Press.
- Bobko, P. (1995). *Correlation and regression: Principles and applications for industrial organizational psychology and management*. New York: McGraw-Hill.
- Bogartz, R. S. (1994). *An introduction to the analysis of variance*. Westport, CT: Praeger.
- Bogdewic, S. P. (1999). Participant observation. In B. J. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (pp. 47–69). Thousand Oaks, CA: Sage.
- Bohman, J. (1991). *New philosophy of social science: Problems of indeterminacy*. Cambridge, UK: Polity.
- Boje, D. M. (2001). *Narrative methods for organizational and communication research*. Thousand Oaks, CA: Sage.
- Bolger, N., Davis, A., & Rafaeli, E. (in press). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*.
- Bollen, K. (1989). *Structural equations with latent variables*. New York: John Wiley.
- Bollen, K. A., & Curran, P. J. (in press). *Latent curve models: A structural equation approach*. New York: John Wiley.
- Bollen, K. A., & Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. In J. Fox & J. S. Long (Eds.), *Modern methods of data analysis* (pp. 257–291). Newbury Park, CA: Sage.
- Bonnie, R. J., & Wallace, R. B. (2002). *Elder mistreatment: Abuse, neglect, and exploitation in an aging America*. Washington, DC: Joseph Henry Press.
- Boomsma, A., Van Duijn, M. A. J., & Snijders, T. A. B. (Eds.). (2001). *Essays on item response theory*. New York: Springer.
- Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling*. New York: Springer-Verlag.
- Borgatta, E. F., & Bohrnstedt, G. W. (1980). Level of measurement: Once over again. *Sociological Methods and Research*, 9, 147–160.
- Borgatta, E. F., & Jackson, D. J. (Eds.). (1980). *Aggregate data: Analysis and interpretation*. Beverly Hills, CA: Sage.
- Bornstein, R. F. (2002). A process dissociation approach to objective-projective test score interrelationships. *Journal of Personality Assessment*, 78, 47–68.
- Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L. (1994). Face validity and fakability of objective and projective measures of dependency. *Journal of Personality Assessment*, 63, 363–386.
- Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.
- Boruch, R. F., & Cecil, J. S. (1979). *Assuring the confidentiality of social research data*. Philadelphia: University of Pennsylvania Press.
- Boster, J. S. (1994). The successive pile sort. *Cultural Anthropology Methods*, 6(2), 7–8.
- Bosworth, M. (1999). *Engendering resistance: Agency and power in women's prisons*. Aldershot, UK: Ashgate/Dartmouth.
- Botteroff, J. L. (1994). Using videotaped recordings in qualitative research. In J. M. Morse (Ed.), *Critical issues in qualitative research* (pp. 224–261). Newbury Park, CA: Sage.
- Boucké, O. F. (1923). The limits of social science: II. *American Journal of Sociology*, 28(4), 443–460.
- Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge, UK: Cambridge University Press.
- Bourdieu, P. (2000). *Pascalian meditations* (R. Nice, Trans.). Cambridge, UK: Polity.
- Bourdieu, P., & Wacquant, L. J. D. (1992). *An invitation to reflexive sociology*. Chicago: University of Chicago Press.
- Bourque, L. B., & Clark, V. A. (1992). *Processing data: The survey example* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–085). Newbury Park, CA: Sage.
- Bourque, L. B., & Fielder, E. P. (2003). *How to conduct self-administered and mail surveys* (2nd ed.). Thousand Oaks, CA: Sage.
- Bourque, L. B., & Russell, L. A. (with Krauss, G. L., Riopelle, D., Goltz, J. D., Greene, M., McAfee, S., & Nathe, S.) (1994, July). *Experiences during and responses*

- to the Loma Prieta earthquake. Oakland: Governor's Office of Emergency Services, State of California.
- Bourque, L. B., Shoaf, K. I., & Nguyen, L. H. (1997). Survey research. *International Journal of Mass Emergencies and Disasters*, 15, 71–101.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society (B)*, 26(2), 211–252.
- Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control* (rev. ed.). San Francisco: Holden-Day.
- Box, G. E. P., & Tiao, G. C. (1965). A change in level of nonstationary time series. *Biometrika*, 52, 181–192.
- Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. New York: Wiley.
- Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70, 70–92.
- Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters*. New York: Wiley.
- Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). New York: Prentice Hall.
- Box-Steffensmeier, J. M., & Jones, B. S. (1997). Time is of the essence: Event history models in political science. *American Journal of Political Science*, 41, 336–383.
- Box-Steffensmeier, J. M., & Zorn, C. J. W. (2001). Duration models and proportional hazards in political science. *American Journal of Political Science*, 45, 951–967.
- Box-Steffensmeier, J. M., & Zorn, C. J. W. (2002). Duration models for repeated events. *The Journal of Politics*, 64(4), 1069–1094.
- Boyd, J. P. (1990). *Social semigroups: A unified theory of scaling and bockmodeling as applied to social networks*. Fairfax, VA: George Mason University Press.
- Boyd, J. P., & Jonas, K. J. (2001). Are social equivalences ever regular? Permutation and exact tests. *Social Networks*, 23, 87–123.
- Bracken, B. (Ed.). (1995). *Handbook of self concept: Developmental, social and clinical consequences*. New York: Wiley.
- Bradley, J. V. (1985). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice Hall.
- Brady, I. (2000). Anthropological poetics. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 949–979). Thousand Oaks, CA: Sage.
- Brady, I. (2003). *The time at Darwin's Reef: Poetic explorations in anthropology and history*. Walnut Creek, CA: AltaMira.
- Bray, J. H., & Maxwell, S. E. (1985). *Multivariate analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–054). Beverly Hills, CA: Sage.
- Breen, R. (1996). *Regression models: Censored, sample selected, or truncated data* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–111). Thousand Oaks, CA: Sage.
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth.
- Brennan, R. L. (1983). *Elements of generalizability*. Iowa City, IA: American College Testing Program.
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Brewer, K. (2002). *Combined survey sampling inference: Weighing Basu's elephants*. London: Arnold.
- Briggs, C. (1986). *Learning how to ask*. Cambridge, UK: Cambridge University Press.
- Broad, W., & Wade, N. (1982). *Betrayers of the truth: Fraud and deceit in the halls of science*. New York: Simon & Schuster.
- Brockmeyer, E., Halstrom, H. L., & Jensen, A. (1960). *The life and works of A. K. Erlang*. Kobenhavn: Akademiet for de Tekniske Videnskaber.
- Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.
- Bronner, S. E., & Kellner, D. M. (1989). *Critical theory and society: A reader*. New York: Routledge.
- Bronshtein, I. N., & Semendyaev, K. A. (1985). *Handbook of mathematics* (K. A. Hirsch, Trans. & Ed.). New York: Van Nostrand Reinhold. (Originally published in 1979)
- Brown, C. (1991). *Ballots of tumult: A portrait of volatility in American voting*. Ann Arbor: University of Michigan Press.
- Brown, C. (1995). *Chaos and catastrophe theories*. Thousand Oaks, CA: Sage.
- Brown, C. (1995). *Serpents in the sand: Essays on the nonlinear nature of politics and human destiny*. Ann Arbor: University of Michigan Press.
- Brown, H. I. (1987). *Observation and objectivity*. New York: Oxford University Press.
- Brown, S. R., & Melamed, L. E. (1998). *Experimental design and analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–074). Thousand Oaks, CA: Sage.
- Brown, S. R., Durning, D. W., & Selden, S. C. (1999). Q methodology. In G. J. Miller & M. L. Whicker (Eds.), *Handbook of research methods in public administration* (pp. 599–637). New York: Dekker.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 12, 296–322.
- Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, UK: Cambridge University Press.
- Browne, M. W. (1984). Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62–83.
- Brownlee, K. (1965). *Statistical theory and methodology in science and engineering*. New York: John Wiley.
- Bruner, J. (1960). *The process of education*. Cambridge, MA: Harvard University Press.
- Bruner, J. (1966). *Toward a theory of instruction*. Cambridge, MA: Harvard University Press.

- Brunk, G. G., Caldeira, G. A., & Lewis-Beck, M. S. (1987). Capitalism, socialism, and democracy: An empirical inquiry. *European Journal of Political Research*, 15, 459–470.
- Brunswik, E. (1943). Organismic achievement and environmental probability. *The Psychological Review*, 50, 255–272.
- Bryant, C. G. A. (1985). *Positivism in social theory and research*. London: Macmillan.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.
- Bryman, A. (2001). *Social research methods*. Oxford, UK: Oxford University Press.
- Bryman, A., & Cramer, D. (1997). *Quantitative data analysis with SPSS for Windows: A guide for social scientists*. New York: Routledge Kegan Paul.
- Bryman, T. S. (1966) *The human perspective in sociology: The methodology of participant observation*. Englewood Cliffs, NJ: Prentice Hall.
- Bulmer, M. (Ed.). (1982). *Social research ethics*. London: Macmillan.
- Bulmer, M., & Solomos, J. (Eds.). (2003). *Researching race and racism*. London: Routledge Kegan Paul.
- Bunge, M. (1959). *Causality: The place of the causal principle in modern science*. Cambridge, MA: Harvard University Press.
- Burawoy, M. (1998). The extended case method. *Sociological Theory*, 16(1), 4–33.
- Burgess, R. (1984). *In the field*. London: Allen & Unwin.
- Burgess, R. G. (Ed.). (1982). *Field research: A sourcebook and field manual*. London: Allen & Unwin.
- Burkhart, R. E. (2000). Economic freedom and democracy: Post-cold war tests. *European Journal of Political Research*, 37, 237–253.
- Burkhart, R. E., and Lewis-Beck, M. S. (1994). Comparative democracy: The economic development thesis. *American Political Science Review*, 88, 903–910.
- Burr, V. (1995). *An introduction to social constructionism*. London: Routledge.
- Burton, M. L. (1975). Dissimilarity measures for unconstrained sorting data. *Multivariate Behavioral Research*, 10, 409–424.
- Bury, M. (2001). Illness narratives: Fact or fiction? *Sociology of Health and Illness*, 23(3), 263–285.
- Buttny, R., & Morris, G. H. (2001). Accounting. In W. P. Robinson & H. Giles (Eds.), *The new handbook of language and social psychology* (pp. 285–301). Chichester, England: Wiley.
- Byrne, D. (1998). *Complexity theory and the social sciences: An introduction*. London: Routledge.
- Byrne, D. (2002). *Interpreting quantitative data*. London: Sage.
- Byrne, D. S., McCarthy, P., Harrison, S., & Keithley, J. (1986). *Housing and health*. Aldershot, UK: Gower.
- Cacioppo, J. T., Petty, R. E., Losch, M. E., & Kim, H. S. (1986). Electromyographic specificity during simple physical and attitudinal tasks: Location and topographical features of integrated EMG responses. *Biological Psychology*, 18, 85–121.
- Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2000). *Handbook of psychophysiology* (2nd ed.). Cambridge, UK: Cambridge University Press.
- Cahill, S. (1987). Children and civility: Ceremonial deviance and the acquisition of ritual competence. *Social Psychology Quarterly*, 50, 312–321.
- Cain, C. (1991). Personal stories: Identity acquisition and self-understanding in Alcoholics Anonymous. *Ethos*, 19, 210–253.
- Cale, G. (2001). *When resistance becomes reproduction: A critical action research study*. Proceedings of the 42nd Adult Education Research Conference. East Lansing: Michigan State University.
- Callendar, J. C., & Osburn, H. G. (1977). A method for maximizing split-half reliability coefficients. *Educational and Psychological Measurement*, 37, 819–825.
- Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. New York: Cambridge University Press.
- Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, 24, 409–429.
- Campbell, D. T. (1988). *Methodology and epistemology for social science: Selected papers* (E. S. Overman, Ed.). Chicago: University of Chicago Press.
- Campbell, D. T. (1991). Methods for the experimenting society. *Evaluation Practice*, 12, 223–260.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.
- Campbell, D. T., Boruch, R. F., Schwartz, R. D., & Steinberg, J. (1977). Confidentiality-preserving modes of access to files and to interfile exchange for useful statistical analysis. *Evaluation Quarterly*, 1, 269–300.
- Cannell, C., Fowler, J., & Marquis, K. (1968). *The influence of interviewer and respondent psychological and behavioral variables on the reporting of household interviews* (Vital and Health Statistics Series 2, No. 6). Washington, DC: National Center for Health Statistics.
- Capdevila, R., & Stainton Rogers, R. (2000). If you go down to the woods today ... : Narratives of Newbury. In H. Addams & J. Proops (Eds.), *Social discourse and environmental policy: An application of Q methodology* (pp. 152–173). Cheltenham, UK: Elgar.
- Carless, S. A. (1998). Assessing the discriminant validity of the transformational leadership behaviour as measured by the MLQ. *Journal of Occupational and Organizational Psychology*, 71, 353–358.
- Carley, K., & Prietula, M. (Eds.). (1994). *Computational organization theory*. Hillsdale, NJ: Lawrence Erlbaum.

- Carlin, B. P., & Louis, T. A. (1998). *Bayes and empirical Bayes methods for data analysis*. London: Chapman & Hall/CRC.
- Carlson, M., & Mulaik, S. A. (1993). Trait ratings from descriptions of behavior as mediated by components of meaning. *Multivariate Behavioral Research*, 28, 111–159.
- Carmines, E. G., & McIver, J. P. (1981). *Unidimensional scaling* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–024). Beverly Hills, CA: Sage.
- Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.
- Carrington, P. J., Scott, J., & Wasserman, S. (Eds.). (2003). *Models and methods in social network analysis*. New York: Cambridge University Press.
- Carroll, J. B. (2002). The five-factor personality model: How complete and satisfactory is it? In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 97–126). Mahwah, NJ: Lawrence Erlbaum.
- Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology*, 31, 607–649.
- Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via a N-way generalisation of Eckart-Young decomposition. *Psychometrika*, 35, 283–299, 310–319.
- Carroll, J. M. (1997). Human-computer interaction: Psychology as a science of design. *Annual Review of Psychology*, 48, 61–83.
- Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford, UK: Oxford University Press.
- Casella, G., & Berger, R. L. (1990). *Statistical inference*. Belmont, CT: Duxbury.
- Casella, G., & Berger, R. L. (2001). *Statistical inference* (2nd ed.). New York: Wadsworth.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276.
- Chadhuri, A., & Mukerjee, R. (1987). *Randomized response: Theory and techniques*. New York: Marcel Dekker.
- Chalmers, A. (1982). *What is this thing called science?* St. Lucia, Australia: University of Queensland Press.
- Chalmers, A. (1999). *What is this thing called science?* (3rd ed.). Buckingham, UK: Open University Press.
- Chamberlayne, P., Bornat, J., & Wengraf, T. (2000). *The turn to biographical methods in social science*. London: Routledge Kegan Paul.
- Chamberlayne, P., Rustin, M., & Wengraf, T. (Eds.). (2002). *Biography and social exclusion in Europe: Experiences and life journeys*. Bristol, UK: Policy Press.
- Charemza, W. W., & Deadman, D. F. (1997). *New directions in econometric practice* (2nd ed.). Cheltenham, UK: Edward Elgar.
- Charlton, J., Patrick, D. L., Matthews, G., & West, P. A. (1981). Spending priorities in Kent: A Delphi study. *Journal of Epidemiology and Community Health*, 35, 288–292.
- Charlton, T., Gunter, B., & Hannan, A. (2002). *Broadcast television effects in a remote community*. Hillsdale, NJ: Lawrence Erlbaum.
- Charmaz, K. (2000). Grounded theory: Constructivist and objectivist methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 509–535). Thousand Oaks, CA: Sage.
- Chatfield, C. (1983). *Statistics for technology* (3rd ed.). London: Chapman & Hall.
- Chatfield, C. (1996). *The analysis of time series: An introduction*. New York: Chapman & Hall.
- Chavez, L., Hubbell, F. A., McMullin, J. M., Martinez, R. G., & Mishra, S. I. (1995). Structure and meaning in models of breast and cervical cancer risk factors: A comparison of perceptions among Latinas, Anglo women and physicians. *Medical Anthropology Quarterly*, 9, 40–74.
- Checkland, P. B. (1981). *Systems thinking, systems practice*. Chichester, UK: Wiley.
- Chell, E. (1998). Critical incident technique. In G. Symon & C. Cassell (Eds.), *Qualitative methods and analysis in organizational research: A practical guide* (pp. 51–72). London: Sage.
- Chell, E., & Baines, S. (1998). Does gender affect business performance? A study of micro-businesses in business services in the U.K. *International Journal of Entrepreneurship and Regional Development*, 10(4), 117–135.
- Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367–405.
- Cherkassky, V., & Mulier, F. (1998). *Learning from data*. New York: Wiley.
- Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide*. New York: Wiley-Interscience.
- Cherryholmes, C. H. (1999). *Reading pragmatism*. New York: Teachers College Press.
- Chiang, A. C. (1984). *Fundamental methods of mathematical economics* (3rd ed.). New York: McGraw-Hill.
- Chiari, G., & Nuzzo, M. L. (1996). Psychological constructivisms: A metatheoretical differentiation. *Journal of Constructivist Psychology*, 9, 163–184.
- Chipman, J. S. (1979). Efficiency of least squares estimation of linear trend when residuals are autocorrelated. *Econometrica*, 47, 115–128.
- Chouliarakis, L., & Fairclough, N. (1999). *Discourse in late modernity*. Edinburgh, UK: Edinburgh University Press.
- Chow, G. (1961). Tests of equality between sets of regression coefficients in linear regression models. *Econometrica*, 28(3), 591–605.
- Chrisman, N. (1997). *Exploring geographic information systems*. New York: Wiley.
- Christians, C. (2000). Ethics and politics in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The handbook of qualitative research* (2nd ed., pp. 133–155). Thousand Oaks, CA: Sage.

- Cicerelli, V. G., & Associates. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development* (2 vols.). Athens: Ohio University and Westinghouse Learning Corp.
- Cicourel, A. V. (1974). *Theory and method in a critique of Argentine fertility*. New York: John Wiley.
- Clark, J. A., & Mishler, E. G. (1992). Attending to patients' stories: Reframing the clinical task. *Sociology of Health and Illness*, 14, 344–370.
- Clarke, H. D., Norpoth, H., & Whiteley, P. F. (1998). It's about time: Modeling political and social dynamics. In E. Scarborough & E. Tanenbaum (Eds.), *Research strategies in the social sciences* (pp. 127–155). Oxford, UK: Oxford University Press.
- Clarke, K. A. (2001). Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science*, 45, 724–744.
- Clayman, S. E., & Maynard, D. (1995). Ethnomethodology and conversation analysis. In P. ten Have & G. Psathas (Eds.), *Situated order: Studies in the social organization of talk and embodied activities* (pp. 1–30). Washington, DC: University Press of America.
- Clayman, S., & Heritage, J. (2002). *The news interview: Journalists and public figures on the air*. Cambridge, UK: Cambridge University Press.
- Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *The British Journal of Mathematical and Statistical Psychology*, 22(1), 49–55.
- Clegg, C. W., & Walsh, S. (1998). Soft systems analysis. In G. Symon & C. M. Cassell (Eds.), *Qualitative methods and analysis in organisational research: A practical guide*. London: Sage.
- Cleveland, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.
- Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart.
- Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596–610.
- Clifford, J., & Marcus, G. E. (Eds.). (1986). *Writing culture: The poetics and politics of ethnography*. Berkeley: University of California Press.
- Clogg, C. C. (1978). Adjustment of rates using multiplicative models. *Demography*, 15, 523–539.
- Clogg, C. C. (1982). Some models for the analysis of association in multiway cross-classifications having ordered categories. *Journal of the American Statistical Association*, 77, 803–815.
- Clogg, C. C. (1982). Using association models in sociological research: Some examples. *American Journal of Sociology*, 88, 114–134.
- Clogg, C. C., & Eliason, S. R. (1988). A flexible procedure for adjusting rates and proportions, including statistical methods for group comparisons. *American Sociological Review*, 53, 267–283.
- Clogg, C. C., & Shihadeh, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.
- Clogg, C. C., Shockey, J. W., & Eliason, S. R. (1990). A general statistical framework for adjustment of rates. *Sociological Methods & Research*, 19, 156–195.
- Clough, T. P. (1998). *The end(s) of ethnography: From realism to social criticism* (2nd ed.). New York: Peter Lang.
- Cobb, G. W. (1998). *Introduction to design and analysis of experiments*. New York: Springer-Verlag.
- Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256–266.
- Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics*, 13(3), 261–281.
- Cochran, W. G. (1965). The planning of observational studies in human populations. *Journal of the Royal Statistical Society, Series A*, 128, 134–155.
- Cochran, W. G. (1953). *Sampling techniques*. New York: Wiley.
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley.
- Cochrane, D., & Orcutt, G. H. (1949). Application of least squares relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, 44, 32–61.
- Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data: Complementary research strategies*. Thousand Oaks, CA: Sage.
- Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the $r_{WG(J)}$ index of agreement. *Psychological Methods*, 6, 297–310.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cohen, J. (1978). Partialled products are interactions; partialled powers are curve components. *Psychological Bulletin*, 85, 858–866.
- Cohen, J. (1988). *Statistical power analysis in the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J. (2001). Smallpox vaccinations: How much protection remains? *Science*, 294, p. 985.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Cole, M. (1996). *Cultural psychology: A once and future discipline*. Cambridge, MA: Belknap Press of Harvard University Press.
- Coleman, C., & Moynihan, J. (1996). *Understanding crime data: Haunted by the dark figure*. Buckingham, UK: Open University Press.
- Collett, D. (1991). *Modelling binary data*. New York: Chapman and Hall.
- Collican, H. (1999). *Research methods and statistics in psychology*. London: Hodder & Staughton.
- Collier, A. (1994). *Critical realism*. London: Verso.
- Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice*. London: Sage.

- Collins, P. H. (1997). Comment on Heckman's "Truth and method: Feminist standpoint theory revisited": Where's the power? *Signs*, 22(21), 375–381.
- Collins, R. (1995). Prediction in macrosociology: The case of the Soviet collapse. *American Journal of Sociology*, 100(6), 1552–1593.
- Colson, F. (1977). *Sociale indicatoren van enkele aspecten van bevolkingsgroei*. Doctoral dissertation, Katholieke Universiteit Leuven, Department of Sociology, Leuven.
- Comstock, D. E. (1994). A method for critical research. In M. Martin & L. C. McIntyre (Eds.), *Readings in the philosophy of social science* (pp. 625–639). Cambridge: MIT Press.
- Confidentiality and Data Access Committee. (1999). *Checklist on disclosure potential of proposed data releases*. Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office. Retrieved from www.fcsm.gov/committees/cdac/checklist_799.doc
- Conger, A. J. (1974). Revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement*, 34, 35–46.
- Connell, R. W. (2002). *Gender*. Cambridge, UK: Polity.
- Conover, W. J. (1980). *Practical nonparametric statistics*. New York: Wiley.
- Conover, W. J. (1998). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.
- Conte, R., & Dellarocas, C. (2001). Social order in info societies: An old challenge for innovation. In R. Conte & C. Dellarocas (Eds.), *Social order in multi-agent systems* (pp. 1–16). Boston: Kluwer Academic.
- Converse, J. (1987). *Survey research in the United States*. Berkeley: University of California Press.
- Converse, J., & Presser, S. (1986). *Survey questions*. Beverly Hills, CA: Sage.
- Converse, P. E. (1964). The nature of belief systems in mass publics. In D. Apter (Ed.), *Ideology and discontent* (pp. 206–261). New York: Free Press.
- Converse, P. E., & Markus, G. B. (1979). Plus ça change . . . : The new CPS Election Study Panel. *American Political Science Review*, 73, 32–49.
- Cook, D. (1996). On the interpretation of regression plots. *Journal of the American Statistical Association*, 91, 983–992.
- Cook, D. R., & Weisberg, S. (1999). *Applied regression including computing and graphics*. New York: Wiley.
- Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally College Publishing.
- Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution.
- Cook, T. D., Campbell, D. T., & Peracchio, L. (1990). Quasi-experimentation. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organisational psychology* (2nd ed., Vol. 1, pp. 491–576). Chicago: Rand McNally.
- Cooke, B., & Kothari, U. (Eds.). (2001). *Participation: The new tyranny?* London: Zed.
- Cooksy, L. J., Gill, P., & Kelly, P. A. (2001). The program logic model as an integrative framework for a multimethod evaluation. *Evaluation and Program Planning*, 24, 119–128.
- Cooley, W., & Lohnes, P. (1971). *Multivariate data analysis*. New York: Wiley.
- Coombs, C. H. (1964). *A theory of data*. New York: Wiley.
- Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.
- Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.
- Copi, I. M., & Cohen, C. (1990). *Introduction to logic* (8th ed.). New York: Macmillan.
- Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88(4), 715–730.
- Cormack, R. (2001). Population size estimation and capture-recapture methods. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 17). Amsterdam: Elsevier.
- Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to algorithms* (2nd ed.). Cambridge: MIT Press.
- Correll, S. (1995). The ethnography of an electronic bar: The lesbian cafe. *Journal of Contemporary Ethnography*, 24(3), 270–298.
- Cortazzi, M. (2001). Narrative analysis in ethnography. In P. Atkinson, A. Coffey, S. Delamont, J. Lofland, & L. Lofland (Eds.), *Handbook of ethnography*. London: Sage.
- Cortazzi, M., Jin, L., Wall, D., & Cavendish, S. (2001). Sharing learning through narrative communication. *International Journal of Language and Communication Disorders*, 36, 252–257.
- Corter, J. E. (1996). *Tree models of similarity and association* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–112). Thousand Oaks, CA: Sage.
- Corti, L. (1993). Using diaries in social research (*Social Research Update*, Iss. 2). Guildford, UK: University of Surrey, Department of Sociology.
- Corti, L., Foster, J., & Thompson, P. (1995). *Archiving qualitative research data* (Social Research Update No. 10). Surrey, UK: Department of Sociology, University of Surrey.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
- Costantino, G., Malgady, R. G., & Vazquez, C. (1981). A comparison of the Murray–TAT and a new Thematic

- Apperception Test for urban Hispanic children. *Hispanic Journal of Behavior Sciences*, 3, 291–300.
- Costigan, P., & Thomson, K. (1992). Issues in the design of CAPI questionnaires for complex surveys. In A. Westlake, R. Banks, C. Payne, & T. Orchard (Eds.), *Survey and statistical computing* (pp. 47–156). London: North Holland.
- Costner, H. L. (1965). Criteria for measures of association. *American Sociological Review*, 30, 341–353.
- Couper, M. P. (2000). WEB surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 464–494.
- Couper, M. P., Baker, R. P., Bethlehem, J., Clark, C. Z. F., Martin, J., Nichols, W. L., et al. (Eds.). (1998). *Computer assisted survey information collection*. New York: John Wiley.
- Courant, Richard. (1988). *Differential and integral calculus* (2 vols.) (E. J. McShane, Trans.). New York: Wiley. (Originally published in 1934)
- Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement*, 61, 229–248.
- Cousins, J. B. (2003). Utilization effects of participatory evaluation. In T. Kellaghan, D. L. Stufflebeam, & L. A. Wingate (Eds.), *International handbook of educational evaluation* (pp. 245–266). Dordrecht, The Netherlands: Kluwer Academic.
- Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. *New Directions in Evaluation*, 80, 5–23.
- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, B*, 34, 187–220.
- Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. New York: Chapman and Hall.
- Cox, R. T. (1990). Probability, frequency, and reasonable expectation. In G. Shafer & J. Pearl (Eds.), *Readings in uncertain reasoning* (pp. 353–365). New York: Morgan Kaufmann. (Original work published 1946)
- Coxon, A. P. M. (1982). *The user's guide to multidimensional scaling*. London: Heinemann Educational.
- Coxon, A. P. M. (1999). *Between the sheets: Sexual diaries and gay men's sex in the era of AIDS*. London: Cassell.
- Coxon, A. P. M. (1999). *Sorting data: Collection and analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–127). Thousand Oaks, CA: Sage.
- Cramer, J. S. (1986). *Econometric applications of maximum likelihood methods*. Cambridge, UK: Cambridge University Press.
- Crano, W. D., & Brewer, M. B. (2002). *Principles and methods of social research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Crapanzano, V. (1980). *Tuhami: Portrait of a Moroccan*. Chicago: University of Chicago Press.
- Crenshaw, K., Gotanda, N., Peller, G., & Thomas, K. (Eds.). (1995). *Critical race theory*. New York: The New York Press.
- Cressey, D. (1950). The criminal violation of financial trust. *American Sociological Review*, 15, 738–743.
- Cressey, D. (1953). *Other people's money*. Glencoe, IL: Free Press.
- Cressie, N. (1993). *Statistics for spatial data*. New York: Wiley.
- Cressie, N., & Read, T. (1984). Multinomial goodness of tests. *Journal of Royal Statistical Society Series B*, 46, 440–464.
- Creswell, J. W. (1995). *Research design: Quantitative and qualitative approaches*. Thousand Oaks, CA: Sage.
- Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. New York: Harcourt Brace Jovanovich.
- Croll, P. (1986). *Systematic classroom observation*. Lewes, UK: Falmer.
- Cromwell, J. B., Labys, W., & Terraza, M. (1994). *Univariate tests for time series models*. Thousand Oaks, CA: Sage.
- Cronbach, L. J. (1943). On estimates of test reliability. *Journal of Educational Psychology*, 34, 485–494.
- Cronbach, L. J. (1946). A case study of the split-half reliability coefficient. *Journal of Educational Psychology*, 37, 473–480.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Cronbach, L. J. (1955). Processes affecting scores on “understanding of others” and “assumed similarity.” *Psychological Bulletin*, 52, 177–193.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: HarperCollins.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.
- Croon, M. A., Bergsma, W., & Hagenaars, J. A. (2000). Analyzing change in categorical variables by generalized log-linear models. *Sociological Methods & Research*, 29, 195–229.
- Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. London: Sage.
- Crowne, D., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: Wiley.
- Cruse, D. A. (1986). *Lexical semantics*. Cambridge, UK: Cambridge University Press.
- Cullen, C. G. (1990). *Matrices and linear transformations* (2nd ed.). New York: Dover.
- Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64, 413–428.
- Czaja, R., Blair, J., & Sebestile, J. P. (1982). Respondent selection in a telephone survey: Comparison of three techniques. *Journal of Marketing Research*, 21, 381–385.
- Czyzewski, M. (1994). Reflexivity of actors versus reflexivity of accounts. *Theory, Culture and Society*, 11, 161–168.
- D'Andrade, R. (1995). *The development of cognitive anthropology*. Cambridge, UK: Cambridge University Press.
- Dale, A., & Marsh, C. (Eds.). (1993). *The 1991 census user's guide*. London: HMSO.

- Dale, A., Arber, S., & Procter, P. (1988). *Doing secondary analysis*. London: Unwin Hyman.
- Dale, A., Fieldhouse, E., & Holdsworth, C. (2000). *Analyzing census microdata*. London: Arnold.
- Dandridge, T. C., Mitroff, I., & Joyce, W. F. (1980). Organizational symbolism: A topic to expand organizational analysis. *Academy of Management Review*, 5, 77–82.
- Daniel, W. W. (1993). *Collecting sensitive data by randomized response: An annotated bibliography* (2nd ed.). Atlanta: Georgia State University Business Press.
- Daniels, H. E. (1944). The relation between measures of correlation in the universe of sample permutations. *Biometrika*, 35, 129–135.
- Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.
- Darnell, A. C. (1995). *A dictionary of econometrics*. Cheltenham, UK: Edward Elgar.
- Darroch, J. N., Lauritzen, S. L., & Speed, T. P. (1980). Markov fields and log-linear models. *Annals of Mathematical Statistics*, 43, 1470–1480.
- David, H. A., & Moeschberg, M. L. (1978). *The theory of competing risks* (Griffin's Statistical Monograph #39). New York: Macmillan.
- David, P. (1985). Clio and the economics of QWERTY. *American Economic Review*, 75, 332–337.
- Davidson, J., Hendry, D. F., Srba, F., & Yeo, S. (1978). Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal*, 88, 661–692.
- Davidson, R., & MacKinnon J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.
- Davis, J. A. (1985). *The logic of causal order* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–055). Beverly Hills, CA: Sage.
- Davis, J. A., & Smith, T. W. (1992). *The NORC General Social Survey: A user's guide*. Newbury Park, CA: Sage.
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.
- Day, N. E. (1969). Estimating the components of a mixture of two normal distributions. *Biometrika*, 56, 463–474.
- De Beauvoir, S. (1970). *The second sex*. New York: Alfred A. Knopf. (Original work published 1949)
- de Leeuw, J., & Kreft, I. G. G. (Eds.). (in press). *Handbook of quantitative multilevel analysis*. Dordrecht, The Netherlands: Kluwer Academic.
- de Leeuw, J., & van der Heijden, P. G. M. (1988). The analysis of time-budgets with a latent time-budget model. In E. Diday (Ed.), *Data analysis and informatics* (Vol. 5, pp. 159–166). Amsterdam: North-Holland.
- de Saussure, F. (1979). *Cours de linguistique générale* (T. de Mauro, Ed.). Paris: Payot.
- de Vaus, D. (Ed.). (2002). *Social surveys* (4 vols.). London: Sage.
- de Vaus, D. A. (1995). *Surveys in social research* (4th ed.). Sydney, Australia: Allen & Unwin.
- de Vaus, D. A. (2001). *Research design in social research*. London: Sage.
- Deacon, D., Bryman, A., & Fenton, N. (1998). Collision or collusion? A discussion of the unplanned triangulation of quantitative and qualitative research methods. *International Journal of Social Research Methodology*, 1, 47–63.
- Deacon, D., Pickering, M., Golding, P., & Murdock, G. (1999). *Researching communications*. London: Arnold.
- DeBoef, S., & Granato, J. (2000). Testing for cointegrating relationships with near-integrated data. *Political Analysis*, 8, 99–117.
- Deckner, D. F., Adamson, L. B., & Bakeman, R. (2003). Rhythm in mother-infant interactions. *Infancy*, 4, 201–217.
- DeGroot M. H., & Schervish, M. J. (2002). *Probability and statistics* (3rd ed.). Reading, MA: Addison-Wesley.
- Dehejia, R. H., & Wahba, S. (1999). Causal effects in non-experimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053–1062.
- Del Monte, K. (2000). Partners in inquiry: Ethical challenges in team research. *International Social Science Review*, 75, 3–14.
- Delanty, G. (1997). *Social science: Beyond constructivism and realism*. Buckingham, UK: Open University Press.
- DeMaris, A. (1992). *Logit modeling*. Newbury Park, CA: Sage.
- Deming, W. E. (1997). *Statistical adjustment of data*. New York: Dover.
- DeNavas-Walt, C., & Cleveland, R. W. (2002). *Money income in the United States* (United States Census Bureau, Current Population Reports, P60–218). Washington, DC: Government Printing Office. Available: www.census.gov/prod/2002pubs/p60-218.pdf
- Denzin, N. (1983). Interpretive interactionism. In G. Morgan (Ed.), *Beyond method: Strategies for social research* (pp. 128–142). Beverly Hills, CA: Sage.
- Denzin, N. (1992). *Symbolic interactionism: The politics of interpretation*. Oxford, UK: Blackwell.
- Denzin, N. K. (1970). *The research act in sociology*. Chicago: Aldine.
- Denzin, N. K. (1989). *Interpretive biography*. Newbury Park, CA: Sage.
- Dezin, N. K., (1989). *Interpretive interactionism*. Newbury Park, CA: Sage.
- Denzin, N. K., (1997). *Interpretive ethnography*. Thousand Oaks, CA: Sage.
- Denzin, N. K., & Lincoln, Y. S. (2000). The discipline and practice of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 1–28). Thousand Oaks, CA: Sage.
- Denzin, N. K., & Lincoln, Y. S. (Eds.). (2000). *Handbook of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.
- Denzin, N. K. & Lincoln, Y. S. (2000). Introduction: The discipline and practice of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 1–28). Thousand Oaks, CA: Sage.
- Derrida, J. (1976). *Of grammatology* (G. Spivak, Trans.). Baltimore, MD: Johns Hopkins University Press.

- Derrida, J. (1995). *The gift of death*. Chicago: The University of Chicago Press.
- Derrida, J., & Ferraris, F. (2001). *A taste for the secret*. Cambridge, UK: Polity.
- DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park, CA: Sage.
- Devore, J. L. (1991). *Probability and statistics for engineering and the sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
- Dexter, L. A. (1970). *Elite and specialized interviewing*. Evanston, IL: Northwest University Press.
- Dey, I. (1993). *Qualitative data analysis*. London: Routledge Kegan Paul.
- Diamond, S. S. (2000). Reference guide on survey research, in *Reference manual on scientific evidence*. (2nd ed., pp. 229–276). Washington, D. C.: Federal Judicial Center.
- Dickens, P. (2000). *Social Darwinism*. Buckingham, UK: Open University Press.
- Diebold, F. X. (2001). *Elements of forecasting* (2nd ed.). Cincinnati, OH: South-Western.
- Dienstbier, R. A. (1989). Arousal and physiological toughness: Implications for mental and physical health. *Psychological Review*, 96, 84–100.
- Diesing, P. (1991). *How does social science work?* Pittsburgh, PA: University of Pittsburgh Press.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41, 417–440.
- Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.
- Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method*. New York: Wiley.
- Dilthey, W. (1985). *Poetry and experience: Selected works* (Vol. 5). Princeton, NJ: Princeton University Press.
- Dion, K. K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, 24, 285–290.
- DiPrete, T. A., & Forristal, J. D. (1994). Multilevel models: Methods and substance. *Annual Review of Sociology*, 20, 331–357.
- Dixon-Woods, M., Fitzpatrick, R., & Roberts, K. (2001). Including qualitative research in systematic reviews: Problems and opportunities. *Journal of Evaluation in Clinical Practice*, 7, 125–133.
- Doksum, K. A., & Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63, 421–434.
- Doreian, P., Batagelj, V., & Ferligoj, A. (1994). Partitioning networks based on generalized concepts of equivalence. *Journal of Mathematical Sociology*, 19, 1–27.
- Douglas, J. D. (1985). *Creative interviewing*. Beverly Hills, CA: Sage.
- Douglass, B., & Moustakas, C. (1985). Heuristic inquiry: The internal search to know. *Journal of Humanistic Psychology*, 25, 39–55.
- Dovidio, J. F., Kawakami, K., & Beach, K. R. (2001). Implicit and explicit attitudes: Examination of the relationship between measures of intergroup bias. In R. Brown & S. L. Gaertner (Eds.), *Blackwell handbook of social psychology: Intergroup processes* (pp. 175–197). Malden, MA: Blackwell.
- Downs, A. (1957). *An economic theory of democracy*. New York: HarperCollins.
- Doyle, P., Martin, B., & Moore, J. (2000). *Improving income measurement*. The Survey of Income Program Participation (SIPP) Methods Panel. Washington, DC: U.S. Bureau of the Census.
- Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.
- Dreher, A. U. (2000). *Foundations for conceptual research in psychoanalysis* (Psychoanalytic Monograph 4). London: Karnac.
- Drew, P., & Heritage, J. (Eds.). (1992). *Talk at work: Interaction in institutional settings*. Cambridge, UK: Cambridge University Press.
- Dryzek, J. S., & Holmes, L. T. (2002). *Post-communist democratization*. Cambridge, UK: Cambridge University Press.
- Dubin, J. A., & Rivers, D. (1989). Selection bias in linear regression, logit and probit models. *Sociological Methods & Research*, 18, 360–390.
- Duda, R., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Wiley.
- Dukes, R. L., Ullman, J. B., & Stein, J. A. (1995). An evaluation of D.A.R.E. (Drug Abuse Resistance Education), using a Solomon four-group design with latent variables. *Evaluation Review*, 19(4), 409–435.
- Duncan, O. D. (1966). Path analysis: Sociological examples. *American Journal of Sociology*, 72, 1–16.
- Duncan, O. D. (1975). *Introduction to structural equation models*. New York: Academic Press.
- Duncan, O. D., Haller, A., & Portes, A. (1968). Peer influence on aspiration: A reinterpretation. *American Journal of Sociology*, 75, 119–137.
- Dupré, J. (2001). *Human nature and the limits of science*. Oxford, UK: Oxford University Press.
- Dupré, J., & Cartwright, N. (1988). Probability and causality: Why Hume and indeterminism don't mix. *Nous*, 22, 521–536.
- Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regressions I. *Biometrika*, 37, 409–428.
- Durkheim, E. (1951). *Suicide*. Glencoe, IL: Free Press.
- Durkheim, E. (1952). *Suicide*. London: Routledge Kegan Paul. (Original work published 1896)
- Durkheim, E. (1964). *The rules of scientific method*. Glencoe, IL: Free Press.
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 276–284.

- Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good but . . . A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, 110, 109–128.
- Easterby-Smith, M., Thorpe, R., & Holman, D. (1996). Using repertory grids in management. *Journal of European Industrial Training*, 20, 1–30.
- Easterby-Smith, M., Thorpe, R., & Lowe, A. (2002). *Management research: An introduction* (2nd ed.). London: Sage.
- Eatwell, J., Milgate, M., & Newman, P. (1987). *The new Palgrave: A dictionary of economics*. London: Macmillan.
- Eckstein, H. (1975). Case study and theory in political science. In F. I. Greenstein & N. W. Polsby (Eds.), *Handbook of political science: Vol. 7, Strategies of inquiry* (pp. 79–137). Reading, MA: Addison-Wesley.
- Eco, U. (1972). Towards a semiotic inquiry into the television message. In *Working papers in cultural studies* (Vol. 3, pp. 103–121). Birmingham: Centre for Contemporary Cultural Studies. (Original work published 1965)
- Eco, U. (1976). *A theory of semiotics*. Bloomington: Indiana University Press.
- Eco, U. (1981). *The role of the reader: Explorations in the semiotics of texts*. London: Hutchinson.
- Economic and Social Research Council Research Centre on Micro-Social Change. (2001, February 28). *British Household Panel Survey* [computer file] (Study Number 4340). Colchester, UK: The Data Archive [distributor].
- Eden, D. (1990). *Pygmalion in management: Productivity as a self-fulfilling prophecy*. Lexington, MA: D. C. Heath.
- Edwards, A. (1948). On Guttman's scale analysis. *Educational and Psychological Measurement*, 8, 313–318.
- Edwards, A. L. (1957). *Techniques of attitude construction*. New York: Appleton-Century-Crofts.
- Edwards, A. W. F. (1972). *Likelihood: An account of the statistical concept of likelihood and its application to scientific inference*. Cambridge, UK: Cambridge University Press.
- Edwards, D. (1997). *Discourse and cognition*. London: Sage.
- Edwards, D. (2000). *Introduction to graphical modelling* (2nd ed.). New York: Springer-Verlag.
- Edwards, D., & Potter, J. (1992). *Discursive psychology*. London: Sage.
- Edwards, D., Ashmore, M., & Potter, J. (1995). Death and furniture: The rhetoric, politics, and theology of bottom line arguments against relativism. *History of the Human Sciences*, 8, 25–49.
- Edwards, W. S., Winn, D. M., Kurlantzick, V. et al. (1994). *Evaluation of National Health Interview Survey diagnostic reporting*. National Center for Health Statistics. Vital Health Stat 2(120).
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1–26.
- Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.
- Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, 40, 1–5.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Egger, M., Davey Smith, G., & Altman, D. G. (2001). *Systematic reviews in health care: Meta-analysis in context* (2nd ed.). London: BMJ Books.
- Egger, M., Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629–634.
- Eibl-Eibesfeldt, I. (1989). *Human ethology*. Hawthorne, NY: Aldine de Gruyter.
- Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, 14, 532–550.
- Eisenhart, M. (1998). On the subject of interpretive reviews. *Review of Educational Research*, 68(4), 391–399.
- Eisner, E. W. (1997). The new frontier in qualitative research methodology. *Qualitative Inquiry*, 3, 259–273.
- Elder, G., Pavalko, E. K., & Clipp, E. C. (1993). *Working with archival lives: Studying lives* (Sage University Papers on Quantitative Applications on the Social Sciences, 07–088). Newbury Park, CA: Sage.
- Eliaison, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. Newbury Park, CA: Sage.
- Ellis, B. D., & Lierse, C. (1994). Dispositional essentialism. *Australasian Journal of Philosophy*, 72(1), 27–45.
- Ellis, C. (1995). *Final negotiations: A story of love, loss, and chronic illness*. Philadelphia: Temple University Press.
- Ellis, C. (2003). *The ethnographic "I": A methodological novel on doing autoethnography*. Walnut Creek, CA: AltaMira.
- Ellis, C. S. (1991). Sociological introspection and emotional experience. *Symbolic Interaction*, 14, 23–50.
- Ellis, C., & Bochner, A. P. (2000). Autoethnography, personal narrative, reflexivity: Researcher as subject. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 733–768). Thousand Oaks, CA: Sage.
- Ellis, C., & Bochner, A. P. (Eds.). (1996). *Composing ethnography: Alternative forms of qualitative writing*. Walnut Creek, CA: AltaMira.
- Elmasri, R. A., & Navathe, S. B. (2001). *Fundamentals of database systems*. New York: Addison-Wesley.
- Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge, UK: Cambridge University Press.
- Elster, J. (1989). *The cement of society: A study of social order*. Cambridge, UK: Cambridge University Press.
- Elster, J. (Ed.). (1986). *Rational choice*. Oxford, UK: Basil Blackwell.
- Ember, C. R., & Ember, M. (2001). *Cross-cultural research methods*. Walnut Creek, CA: AltaMira.
- Embre, L., Behnke, E. A., Carr, D., Evans, J. C., Huertas-Jourda, J., Kockelmans, J. J., et al. (Eds.). (1996). *The encyclopedia of phenomenology*. Dordrecht, The Netherlands: Kluwer.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

- Emerson, R. M., & Pollner, M. (1988). On the use of members' responses to researchers' accounts. *Human Organization*, 47, 189–198.
- Emerson, R. M., Fretz, R. I., & Shaw, L. L. (1995). *Writing ethnographic fieldnotes*. Chicago: University of Chicago Press.
- Emirbayer, M., & Mische, A. (1998). What is agency? *American Journal of Sociology*, 103(4), 962–1023.
- Emmison, M., & Smith, P. (2000). *Researching the visual*. London: Sage.
- Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica*, 55, 251–276.
- Engle, R. F., Granger, C. W. J. (1991). *Long run relationships: Readings in cointegration*. New York: Oxford University Press.
- Engle, R. F., Hendry, D. F., & Richard, J. F. (1983). Exogeneity. *Econometrica*, 51, 277–304.
- Epley, N., & Huff, C. (1998). Suspicion, affective response, and educational benefit of deception in psychology research. *Personality and Social Psychology Bulletin*, 24, 759–768.
- Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up*. Cambridge: MIT Press.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge: MIT Press.
- Erikson, E. H. (1975). *Life history and the historical moment*. New York: Norton.
- Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Inter-generational mobility in three Western European societies. *British Journal of Sociology*, 30.
- Erlang, A. K. (1935). *Fircefrede logaritmetavler og andre regnetavler til brug ved undervisning og i praksis*. Kobenhavn: G.E.C. Gads.
- Ermarth, M. (1978). *Wilhelm Dilthey: The critique of historical reason*. Chicago: University of Chicago Press.
- Escofier, B., & Pagès, J. (1988). *Analyses factorielles multiples* [Multiple factor analyses]. Paris: Dunod.
- Essed, Ph., & Goldberg, T. D. (Eds.). (2002). *Race critical theories: Text and context*. Malden, MA: Blackwell.
- Eubank, R. L. (1988). Quantiles. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 7, pp. 424–432). New York: Wiley.
- European Social Survey Central Co-ordinating Team. (2001). *European Social Survey (ESS): Specification for participating countries*. London: Author.
- Eurostat. (1998). *Labour force survey: Methods and definitions* (1998 ed.). Luxembourg: Author.
- Evans, J. L. (2000). *Early childhood counts*. Washington, DC: World Bank.
- Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical distributions* (3rd ed.). New York: Wiley.
- Everitt, B. S. (1992). *The analysis of contingency tables*. London: Chapman & Hall.
- Everitt, B., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). Oxford, UK: Oxford University Press.
- Exner, J. E. (2002). *The Rorschach: A comprehensive system* (Vol. 1). New York: John Wiley.
- Fairclough, N. (2000). *New labour, new language?* London: Routledge Kegan Paul.
- Fairclough, N., & Wodak, R. (1997). Critical discourse analysis. In T. van Dijk (Ed.), *Discourse as social interaction*. Thousand Oaks, CA: Sage.
- Fals-Borda, O., & Anisur Rahman, M. (Eds.). (1991). *Action and knowledge: Breaking the monopoly with participatory action-research*. New York: Apex.
- Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman & Hall.
- Faugier, J., & Sargeant, M. (1997). Sampling hard to reach populations. *Journal of Advanced Nursing*, 26, 790–797.
- Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, 54, 297–327.
- Featherman, D. L., & Hauser, R. M. (1978). *Opportunity and change*. New York: Academic Press.
- Featherstone, M. (1988). In pursuit of the postmodernism: An introduction. *Theory, Culture & Society*, 5, 195–215.
- Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf & Härtel.
- Fechner, G. T. (1877). *In sachen der Psychophysik*. Leipzig: Breitkopf & Härtel.
- Federal Committee on Statistical Methodology. (1994). *Report on statistical disclosure limitation methodology* (Statistical Policy Working Paper #22). Prepared by the Subcommittee on Disclosure Limitation Methodology. Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office. Retrieved from www.fcsm.gov/working-papers/wp22.html
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.
- Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed.). New York: John Wiley.
- Ferrell, J., & Hamm, M. S. (1998). True confessions: Crime, deviance, and field research. In J. Ferrell & M. Hamm (Eds.), *Ethnography at the edge: Crime, deviance and field research* (pp. 2–19). Boston: Northeastern University Press.
- Fetterman, D. M. (1998). *Ethnography: Step by step* (2nd ed.). Thousand Oaks, CA: Sage.
- Fetterman, D. M. (2002). Web surveys to digital movies: Technological tools of the trade. *Educational Researcher*, 31(6), 29–37.
- Feyerabend, P. (1978). *Against method*. London: Verso.
- Feynman, R. P., & Weinberg, S. (1987). *Elementary particles and the laws of physics: The 1986 Dirac Memorial Lectures*. Cambridge, UK: Cambridge University Press.
- Field, A. (1998). A bluffer's guide to . . . sphericity. *British Psychological Society: Mathematical, Statistical & Computing Newsletter*, 6, 13–22.
- Fielding, J. (1993). Coding and managing data. In N. Gilbert (Ed.), *Researching social life* (pp. 218–238). Thousand Oaks, CA: Sage.

- Fielding, N. G., & Lee, R. M. (1998). *Computer analysis and qualitative research*. London: Sage.
- Fienberg, S. E. (1977). *The analysis of cross-classified categorical data*. Cambridge: MIT Press.
- Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge: MIT Press.
- Fillmore, C. (1975). An alternative to checklist theories of meaning. In C. Cogen, H. Thompson, G. Thurgood, K. Whistler, & J. Wright (Eds.), *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society* (pp. 123–131). Berkeley, CA: Berkeley Linguistics Society.
- Filmer, P. (1998). Analysing literary texts. In C. Seale (Ed.), *Researching society and culture*. London: Sage.
- Fink, A., & Kosecoff, J. (1988). *How to conduct surveys: A step-by-step guide*. Newbury Park, CA: Sage.
- Firebaugh, G. (2001). Ecological fallacy. *International encyclopedia for the social and behavioral sciences* (Vol. 6, pp. 4023–4026). Oxford, UK: Pergamon.
- Fisher, C. B., & Wallace, S. A. (2000). Through the community looking glass: Reevaluating the ethical and policy implications of research on adolescent risk and psychopathology. *Ethics & Behavior*, 10(2), 99–118.
- Fisher, R. A. (1925). *Statistical methods for research workers* (1st ed.). Edinburgh, UK: Oliver & Boyd.
- Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.
- Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London, Series A*, 144, 285–307.
- Fisher, R. A. (1935). *The design of experiments* (1st ed.). Edinburgh, UK: Oliver and Boyd.
- Fisher, R. A. (1971). *Design of experiments*. New York: Hafner Press. (Original work published 1935).
- Fisher, R. A., & Yates, F. (1934). The six by six Latin squares. *Proceedings of the Cambridge Philosophical Society*, 30, 492–507.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, 51(4), 327–358.
- Flick, U. (2002). *An introduction to qualitative research* (2nd ed.). London: Sage.
- Foley, D. E. (1990). *Learning capitalist culture: Deep in the heart of Texas*. Philadelphia: University of Pennsylvania Press.
- Fontana, A., & Frey, J. H. (1994). Interviewing: The art of science. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 361–376). Thousand Oaks, CA: Sage.
- Fontana, A., & Frey, J. H. (2000). The interview: From structured questions to negotiated text. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 645–672). Thousand Oaks, CA: Sage.
- Forrester, J. W. (1961). *Industrial dynamics*. Cambridge: MIT Press.
- Forster, E., & McCleery, A. (1999). Computer assisted personal interviewing: A method of capturing sensitive information. *IASSIST Quarterly*, 23(2), 26–38.
- Foster, J., & Sheppard, J. (1995). *British archives: A guide to archival resources in the United Kingdom* (3rd ed.). London: Macmillan.
- Foucault, M. (1985). *Power/knowledge: Selected interviews and other writings: 1972–1977* (C. Gordon, Ed.). New York: Pantheon.
- Foucault, M. (1991). Questions of method. In G. Burchell, C. Gordon, & P. Miller (Eds.), *The Foucault effect: Studies in governmentality* (pp. 73–86). London: Harvester Wheatsheaf.
- Foucault, M. (1991). *Remarks on Marx: Conversations with Duccio Trombadori* (R. J. Goldstein & J. Cascaito, Trans.). New York: Semiotext(e).
- Foucault, M. (1995). *Discipline and punish: The birth of the prison*. New York: Vintage Books.
- Foucault, M. (1998). *The will to knowledge: The history of sexuality* (Vol. 1). London: Penguin.
- Foucault, M. (2002). *Archeology of knowledge*. London: Routledge.
- Fowler, F. J., Jr. (1988). *Survey research methods* (Rev. ed.). Newbury Park, CA: Sage.
- Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: Sage.
- Fowler, F. J., Jr. (2002). *Survey research methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized survey interviewing*. Newbury Park, CA: Sage.
- Fowler, R. L. (1985). Point estimates and confidence intervals in measures of association. *Psychological Bulletin*, 98, 160–165.
- Fowler, R. L. (1985). Testing for substantive significance in applied research by specifying non-zero effect null-hypotheses. *Journal of Applied Psychology*, 70, 215–218.
- Fox, J. (1991). *Regression diagnostics* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–079). Newbury Park, CA: Sage.
- Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
- Fox, J. (2000). *Nonparametric simple regression: Smoothing scatterplots*. Thousand Oaks, CA: Sage.
- Fox, J. (2000). *Multiple and generalized nonparametric regression*. Thousand Oaks, CA: Sage.
- Fox, J. (2002). *An R and S-PLUS companion to applied regression*. Thousand Oaks, CA: Sage.
- Fox, J. A., & Tracy, P. E. (1986). *Randomized response: A method for sensitive surveys*. Beverly Hills, CA: Sage.
- Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87, 178–183.
- Fox, W. (1998). *Social statistics* (3rd ed.). Bellevue, WA: MicroCase.
- Frank, I. E., & Friedman, J. H. (1993). A statistical view of chemometrics regression tools. *Technometrics*, 35, 109–148.
- Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81, 832–842.

- Frankfort-Nachmias, C., & Nachmias, D. (1996). *Research methods in the social sciences* (5th ed.). New York: St. Martin's.
- Frankfort-Nachmias, C., & Nachmias, D. (2000). *Research methods in the social sciences* (6th ed.). New York: Worth.
- Fransella, F., & Bannister, D. (1977). *A manual for repertory grid technique*. London: Academic Press.
- Franses, P. H. (1998). *Time series models for business and economic forecasting*. Cambridge, UK: Cambridge University Press.
- Franzese, R. (2002). *Macroeconomic policy of developed democracies*. Cambridge, UK: Cambridge University Press.
- Franzosi, R. (2003). *From words to numbers*. Cambridge, UK: Cambridge University Press.
- Fraser, N. (1997). *Justice interruptus: Critical reflections on the 'postsocialist' condition*. London: Routledge Kegan Paul.
- Frederick, B. N. (1999). Partitioning variance in the multivariate case: A step-by-step guide to canonical commonality analysis. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 305–318). Stamford, CT: JAI.
- Frederick, R. I., & Crosby, R. D. (2000). Development and validation of the Validity Indicator Profile. *Law and Human Behavior*, 24, 59–82.
- Freedman, D. A. (2001). Ecological inference and the ecological fallacy. *International encyclopedia for the social and behavioral sciences* (Vol. 6, pp. 4027–4030). Oxford, UK: Pergamon.
- Freedman, D. A., & Wachter, K. W. (2001). *On the likelihood of improving the accuracy of the census through statistical adjustment* (Tech. Rep. 612). Berkeley: University of California, Department of Statistics.
- Freedman, D. A., Klein, S. P., Ostland, M., & Roberts, M. R. (1998). Review of "A solution to the ecological inference problem." *Journal of the American Statistical Association*, 93, 1518–1522. (Discussion appears in Vol. 94, pp. 352–357)
- Freedman, D. A., Pisani, R., & Purves, R. A. (1998). *Statistics*. 3rd ed. New York: W. W. Norton, Inc.
- Freeman, D. (1983). *Margaret Mead and Samoa*. Cambridge, MA: Harvard University Press.
- Freeman, J. F. (1983). Granger causality and the time series analysis of political relationships. *American Journal of Political Science*, 27, 325–355.
- Frey, J. H. (1989). *Survey research by telephone* (2nd ed.). Newbury Park, CA: Sage.
- Frey, J. H., & Fontana, A. (1991). The group interview in social research. *The Social Science Journal*, 28, 175–187.
- Frey, J. H., & Oishi, S. M. (1995). *How to conduct interviews by telephone and in person*. Thousand Oaks, CA: Sage.
- Fricker, R. D., & Schonlau, M. (2002). Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods*, 14(4), 347–365.
- Friedkin, N. (1998). *A structural theory of social influence*. New York: Cambridge University Press.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 675–701.
- Friedman, M. (1962). The interpolation of time series by related series. *Journal of the American Statistical Association*, 57, 729–757.
- Friedman, M. (1991). The re-evaluation of logical positivism. *Journal of Philosophy*, 88(10), 505–519.
- Friedricks, R. (1970). *A sociology of sociology*. New York: Free Press.
- Friendly, M. (1992). Mosaic displays for loglinear models. In American Statistical Association (Ed.), *Proceedings of the Statistical Graphics Section* (pp. 61–68). Alexandria, VA: American Statistical Association.
- Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89, 190–200.
- Friendly, M. (1998). Mosaic displays. In S. Kotz, C. Reed, & D. L. Banks (Eds.), *Encyclopedia of statistical sciences* (Vol. 2, pp. 411–416). New York: John Wiley.
- Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3), 373–395.
- Friendly, M. (2000). *Visualizing categorical data*. Cary, NC: SAS Institute.
- Friendly, M. (2002). A brief history of the mosaic display. *Journal of Computational and Graphical Statistics*, 11(1), 89–107.
- Fu, V., Mare, R., & Winship, C. (2003). Sample selection bias models. In M. A. Hardy & A. Bryman (Eds.), *Handbook of data analysis*. London: Sage.
- Fuller, D. (1999). Part of the action, or "going native"? Learning to cope with the politics of integration. *Area*, 31(3), 221–227.
- Fuller, W. A. (1987). *Measurement error models*. New York: John Wiley.
- Fuss, D. (1989). *Essentially speaking*. New York: Routledge.
- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58, 453–467.
- Gadamer, H.-G. (1975). *Truth and method*. New York: Seabury.
- Gadamer, H.-G. (1989). *Truth and method* (rev. 2nd ed.). New York: Crossroad.
- Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 87, 564–567.
- Gallmeier, C. P. (1991). Leaving, revisiting, and staying in touch: Neglected issues in field research. In W. B. Shaffir & R. A. Stebbins (Eds.), *Fieldwork experience: Qualitative approaches to social research* (pp. 224–231). London: Sage.
- Galton, F. (1886). Regression toward mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, 246–263.
- Galton, F. (1889) *Natural inheritance*. London: Macmillan.

- Galton, M., Simon, B., & Croll, P. (1980). *Inside the primary classroom*. London: Routledge and Kegan Paul.
- Galunic, D. C., & Eisenhardt, K. M. (2001). Architectural innovation and modular corporate forms. *Academy of Management Journal*, 44(6), 1229–1249.
- Game, A. (1991). *Undoing the social: Towards a deconstructive sociology*. Buckingham, UK: Open University Press.
- Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice Hall.
- Garfinkel, H. (1974). On the origins of the term “ethnomethodology.” In R. Turner (Ed.), *Ethnomethodology* (pp. 15–18). Harmondsworth, UK: Penguin.
- Garfinkel, H. (2002). *Ethnomethodology's program: Working out Durkheim's aphorism*. Blue Ridge Summit, PA: Rowman and Littlefield.
- Garman, M. (1990). *Psycholinguistics*. Cambridge, UK: Cambridge University Press.
- Garson, G. D. (2002, June). *Statnotes: An online text-book* [Online]. Retrieved from <http://www2.chass.ncsu.edu/garson/pa765/statnote.htm>.
- Gaventa, J. (1980). *Power and powerlessness: Quiescence and rebellion in an Appalachian valley*. Chicago: University of Chicago Press.
- Gee, J. P. (1991). A linguistic approach to narrative. *Journal of Narrative and Life History*, 1, 15–39.
- Geertz, C. (1973). *The interpretation of cultures*. New York: Basic Books.
- Geertz, C. (1983). *Local knowledge*. New York: Basic Books.
- Geertz, C. (1988). *Works and lives: The anthropologist as author*. Stanford, CA: Stanford University Press.
- Geertz, C. (2000). *The interpretation of cultures: Selected essays*. New York: Basic Books.
- Geladi, P., & Kowalski, B. (1986). Partial least square regression: A tutorial. *Analytica Chimica Acta*, 35, 1–17.
- Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.
- Gensler, H. J. (2002). *Introduction to logic*. London: Routledge Kegan Paul.
- Gentle, J. (2003). *Computational statistics*. New York: Springer-Verlag.
- George, A. L. (1979). Case studies and theory development: The method of structured, focused comparison. In P. G. Lauren (Ed.), *Diplomacy: New approaches in history, theory, and policy* (pp. 43–68). New York: Free Press.
- George, A. L., & McKeown, T. J. (1985). Case studies and theories of organizational decision making. *Advances in Information Processing in Organizations*, 2, 21–58.
- George, A. L., & Smoke, R. (1974). *Deterrence in American foreign policy: Theory and practice*. New York: Columbia University Press.
- Georges, R. A., & Jones, M. O. (1995). *Folkloristics: An introduction*. Bloomington: Indiana University Press.
- Gephart, R. P. (1986). Deconstructing the defence for quantification in social science: A content analysis of journal articles on the parametric strategy. *Qualitative Sociology*, 9, 126–144.
- Gephart, R. P. (1988). *Ethnostatistics: Qualitative foundations for quantitative research*. Newbury Park, CA: Sage.
- Gephart, R. P. (1997). Hazardous measures: An interpretive textual analysis of quantitative sensemaking during crises. *Journal of Organizational Behavior*, 18, 583–622.
- Gerber, A. S., & Green, D. P. (2000). The effects of personal canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *The American Political Science Review*, 94, 653–664.
- Gergen, K. J. (1994). *Realities and relationships*. Cambridge, MA: Harvard University Press.
- Gergen, K. J. (1999). *An invitation to social construction*. London: Sage.
- Gerhardt, U. (1985). Erzählenden und Hypothesenkonstruktion. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 37, 230–256.
- Gerhardt, U. (1986). *Patientenkarrieren: Eine idealtypen-analytische Studie*. Frankfurt: Suhrkamp.
- Gerhardt, U. (1994). The use of Weberian ideal-type methodology in qualitative data interpretation: An outline for ideal-type analysis. *BMS Bulletin de Methodologie Sociologique* (International Sociological Association, Research Committee 33), 45, 76–126.
- Gerhardt, U. (1999). *Herz und Handlungsrationaltät: Eine idealtypen-analytische Studie*. Frankfurt: Suhrkamp.
- Gershuny, J. (2000). *Changing times: Work and leisure in post-industrial societies*. Oxford, UK: Oxford University Press.
- Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: W. H. Freeman.
- Gibbons, J. D. (1988). Truncated data. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 9, p. 355). New York: Wiley.
- Gibbons, J. D. (1993). *Nonparametric statistics: An introduction*. Newbury Park, CA: Sage.
- Gibbons, J. D. (1997). *Nonparametric methods for quantitative analysis* (3rd ed.). Columbus, OH: American Sciences Press.
- Gibbons, J. D., & Chakraborti, S. (1992). *Nonparametric statistical inference* (3rd ed.). New York: Marcel Dekker.
- Gibson, N., Gibson, G., & Macaulay, A. C. (2001). Community-based research: Negotiating agendas and evaluating outcomes. In J. Morse, J. Swanson, & A. J. Kuzel (Eds.), *The nature of qualitative evidence* (pp. 160–182). Thousand Oaks, CA: Sage.
- Giddens, A. (1976). *New rules of sociological method*. London: Hutchinson.

- Giddens, A. (1979). *Central problems in social theory*. Berkeley: University of California Press.
- Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Cambridge, UK: Polity.
- Giele, J. Z., & Elder, G. H. (1998). *Methods of life course research: Qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.
- Gifi, A. (1990). *Nonlinear multivariate analysis*. New York: Wiley.
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650–669.
- Gilbert, G. N., & Mulkay, M. (1984). *Opening Pandora's box: A sociological analysis of scientists' discourse*. Cambridge, UK: Cambridge University Press.
- Gilbert, N., & Troitzsch, K. G. (1999). *Simulation for the social scientist*. Milton Keynes, UK: Open University Press.
- Gill, J. (2000). *Generalized linear models: A unified approach*. Thousand Oaks, CA: Sage.
- Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. London: Chapman & Hall.
- Gintis, H. (2000). *Game theory evolving*. Princeton, NJ: Princeton University Press.
- Girden, E. R. (1992). *ANOVA: Repeated measures* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–084). Newbury Park, CA: Sage.
- Glaser, B. (1992). *Emergence vs. forcing: Basics of grounded theory analysis*. Mill Valley, CA: Sociology Press.
- Glaser, B. G. (1978). *Theoretical sensitivity*. Mill Valley, CA: Sociology Press.
- Glaser, B. G. (2002). Grounded theory and gender relevance. *Health Care for Women International*, 23, 786–793.
- Glaser, B. G., & Strauss, A. L. (1967). *Discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.
- Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the analysis of variance and covariance. *Review of Educational Research*, 42, 237–288.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Research*, 5, 3–8.
- Glass, G. V. (1977). Integrating findings. *Review of Research in Education*, 5, 351–379.
- Glass, G. V., & Smith, M. L. (1978). Meta-analysis of research on the relationship of class size and achievement. *Educational Evaluation and Policy Analysis*, 1, 2–16.
- Glenn, N. D. (2003). Distinguishing age, period, and cohort effects. In J. Mortimer & M. Shanahan (Eds.), *Handbook of the life course* (pp. 465–476). New York: Kluwer Academic/Plenum.
- Gluck, S. B., & Patai, D. (Eds.). (1991). *Women's words: The feminist practice of oral history*. London: Routledge.
- Glymour, C. N. (1980). *Theory and evidence*. Princeton, NJ: Princeton University Press.
- Goffman, E. (1959). *The presentation of self in everyday life*. Garden City, NY: Doubleday.
- Goffman, E. (1983). The interaction order. *American Sociological Review*, 48, 1–17.
- Gold, R. L. (1958). Roles in sociological field observation. *Social Forces*, 36, 217–223.
- Gold, R. L. (1969). Roles in sociological field observations. In G. J. McCall & J. L. Simmons (Eds.), *Issues in participant observation* (pp. 30–38). Reading, MA: Addison-Wesley.
- Goldberger, A. S. (1964). *Econometric theory*. New York: John Wiley & Sons.
- Goldmann, L. (1981). *Method in the sociology of literature* (W. Boelhower, Trans. & Ed.). Oxford, UK: Basil Blackwell.
- Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffin.
- Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.
- Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Hodder Arnold.
- Golub, G. H., & Van Loan, G. F. (1996). *Matrix computations* (3rd ed.). Baltimore: Johns Hopkins University Press.
- Gomm, R., Hammersley, M., & Foster, P. (2000). Case study and generalisation. In R. Gomm, M. Hammersley, & P. Foster (Eds.), *Case study method: Key issues, key texts*. London: Sage.
- Gomm, R., Hammersley, M., & Foster, P. (Eds.). (2000). *Case study method*. London: Sage.
- Good, I. J. (1988). The interface between statistics and philosophy of science. *Statistical Science*, 3, 386–397.
- Good, P. (1994). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. New York: Springer.
- Goode, W. J. (1978). *The celebration of heroes: Prestige as a control system*. Berkeley: University of California Press.
- Goodenough, W. (1957). Cultural anthropology and linguistics. In P. L. Garvin (Ed.), *Report of the 7th Annual Roundtable on Linguistics and Language Study* (pp. 167–173). Washington, DC: Georgetown University Press.
- Goodman, A., Johnson, P., & Webb, S. (1997). *Inequality in the UK*. Oxford, UK: Oxford University Press.
- Goodman, L. (1953). Ecological regression and the behavior of individuals. *American Sociological Review*, 18, 663–664.
- Goodman, L. A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56, 841–868.
- Goodman, L. A. (1973). Causal analysis of data from panel studies and other kinds of surveys. *American Journal of Sociology*, 78, 1135–1191.
- Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to

- others: A modified path analysis approach. *Biometrika*, 60, 179–192.
- Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable: I. A modified latent structure approach. *American Journal of Sociology*, 79, 1179–1259.
- Goodman, L. A. (1978). *Analyzing qualitative/categorical data*. Cambridge, MA: Abt.
- Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74, 537–552.
- Goodman, L. A. (1984). *The analysis of cross-classified data having ordered categories*. Cambridge, MA: Harvard University Press.
- Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review*, 54, 243–309.
- Goodman, L. A., & Hout, M. (1998). Understanding the Goodman-Hout approach to the analysis of differences in association and some related comments. In A. E. Raftery (Ed.), *Sociological methodology* (pp. 249–261). Washington, DC: American Sociological Association.
- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association*, 49, 732–764.
- Goodman, S. N. (1999). Toward evidence-based medical statistics: 1. The *p* value fallacy. *Annals of Internal Medicine*, 130, 995–1004.
- Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.
- Gopaul-McNicol, S.-A., & Armour-Thomas, E. (2002). *Assessment and culture: Psychological tests with minority populations*. San Diego: Academic Press.
- Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. Cambridge, UK: Cambridge University Press.
- Gottschalk, L., Kluckhohn, C., & Angell, R. (1945). *The use of personal documents in history, anthropology, and sociology*. New York: Social Science Research Council.
- Gouldner, A. V. (1973). *For sociology*. Harmondsworth, UK: Penguin.
- Gouldner, A. W. (1962). Anti-minotaur: The myth of a value-free sociology. *Social Problems*, 9, 199–213.
- Gower, J. C., & Hand, D. J. (1996). *Biplots*. London: Chapman & Hall.
- Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 5, 5–48.
- Graham, J. M., Guthrie, A. C., & Thompson, B. (2003). Consequences of not interpreting structure coefficients in published CFA research: A reminder. *Structural Equation Modeling*, 10, 142–153.
- Grammer, K., Fink, B., & Renninger, L. (2002). Dynamic systems and inferential information processing in human communication. *Neuroendocrinology Letters*, 23(Suppl. 4), 15–22.
- Gramsci, A. (1971). *Selections from the prison notebooks*. London: Lawrence and Wishart.
- Granger, C. W. J. (1964). *Spectral analysis of economic time series*. Princeton, NJ: Princeton University Press.
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 24–36.
- Granger, C. W. J. (1981). Some properties of time series data and their use in econometric models specification. *Journal of Econometrics*, 16, 121–130.
- Granger, C. W. J. (1983). *Co-integrated variables and error-correcting models*. Discussion Paper No. 1983–13, University of California, San Diego.
- Granger, C. W. J. (1989). *Forecasting in business and economics*. Boston: Academic Press.
- Granger, C. W. J., & Newbold, P. (1986). *Forecasting economic time series* (2nd ed.). New York: Academic Press.
- Gravetter, F. J., & Wallnau, L. B. (2000). *Statistics for the behavioral sciences* (5th ed.). Belmont, CA: Wadsworth.
- Gray, J., & Reuter, A. (1992). *Transaction processing: Concepts and techniques*. San Francisco: Morgan Kaufmann.
- Green, D. P., & Gerber, A. S. (2002). Reclaiming the experimental tradition in political science. In I. Katznelson & H. V. Milner (Eds.), *Political science: State of the discipline* (3rd ed., pp. 805–832). New York: W.W. Norton.
- Green, D. P., Gerber, A. S., & De Boef, S. L. (1999). Tracking opinion over time: A method for reducing sampling error. *Public Opinion Quarterly*, 63, 178–192.
- Green, P. (1978). *Analyzing multivariate data*. Hinsdale, IL: Dryden.
- Green, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River: Prentice Hall.
- Greenacre, M. J. (1984). *Theory and method of correspondence analysis*. London: Academic Press.
- Greenacre, M. J. (1993). *Correspondence analysis in practice*. London: Academic Press.
- Greenacre, M. J., & Blasius, J. (1994). *Correspondence analysis in the social sciences*. London: Academic Press.
- Greenbaum, T. L. (1998). *The handbook for focus group research*. London: Sage.
- Greene, R. L. (2000). *The MMPI-2: An interpretive manual*. Boston: Allyn & Bacon.
- Greene, W. H. (1993). *Econometric analysis* (2nd ed.). New York: Macmillan.
- Greene, W. H. (1997). *Econometric analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Greene, W. H. (2000). *Econometric analysis* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Greenfield, P. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist*, 52(10), 1115–1124.

- Greenhouse, G. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 55, 431–433.
- Greenland, S. (2003). Quantifying biases in causal models. *Epidemiology*, 14, 300–306.
- Greenland, S., & Brumback, B. A. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology*, 31, 1030–1037.
- Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15, 413–419.
- Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14, 29–46.
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109, 3–25.
- Greenwood, D., & Levin, M. (1998). *Introduction to action research: Social research for social change*. Thousand Oaks, CA: Sage.
- Greimas, A. J., & Courtés, J. (1982). *Semiotics and language: An analytical dictionary*. Bloomington: Indiana University Press.
- Griffin, L., Caplinger, C., Lively, K., Malcom, N. L., McDaniel, D., & Nelsen, C. (1997). Comparative-historical analysis and scientific inference: Disfranchisement in the U.S. South as a test case. *Historical Methods*, 30, 13–27.
- Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica*, 25, 501–522.
- Grinyer, A. (2002). The anonymity of research participants: Assumptions, ethics, and practicalities. *Social Research Update*, 36, 1–4.
- Grofman, B., & Davidson, C. (1992). *Controversies in minority voting: The voting rights act in perspective*. Washington, DC: Brookings Institution.
- Gross, D., & Harris, C. M. (1998). *Fundamentals of queueing theory* (3rd ed.). New York: Wiley.
- Groves, E. R., & Ogburn, W. F. (1928). *American marriage and family relationships*. New York: Holt.
- Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
- Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.
- Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.). (2001). *Survey nonresponse*. New York: Wiley.
- Guba, E. G. (1978). *Toward a methodology of naturalistic inquiry in educational evaluation*. CSE Monograph Series in Evaluation, No. 8. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.
- Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communications and Technology Journal*, 29, 75–92.
- Guba, E. G., & Lincoln, Y. S. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.
- Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.
- Guba, E., & Lincoln, Y. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105–117). Thousand Oaks, CA: Sage.
- Gubrium, J. F., & Buckholdt, D. R. (1979). Production of hard data in human service organizations. *Pacific Sociological Review*, 22, 115–136.
- Gubrium, J. F., & Holstein, J. A. (1997). *The new language of qualitative method*. New York: Oxford University Press.
- Gubrium, J. F., & Holstein, J. A. (2002). From the individual interview to the interview society. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research: Context and method* (pp. 3–32). Thousand Oaks, CA: Sage.
- Gujarati, D. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.
- Gujarati, D. (2002). *Basic econometrics* (4th ed.). New York: McGraw-Hill.
- Gulliksen, H. (1950). *The theory of mental tests*. New York: Wiley.
- Guttman, L. (1941). An outline for the statistical theory of prediction. In P. Horst (with collaboration of P. Wallin & L. Guttman) (Ed.), *The prediction of personal adjustment* (Bulletin 48, Supplementary Study B-1, pp. 253–318). New York: Social Science Research Council.
- Guttman, L. (1944). A basis for scaling quantitative data. *American Sociological Review*, 9, 139–150.
- Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika*, 18, 277–296.
- Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, 19, 149–161.
- Guttman, L. (1956). “Best possible” systematic estimates of communalities. *Psychometrika*, 21, 273–285.
- Haack, S. (1993). *Evidence and inquiry: Toward a reconstruction of epistemology*. Oxford, UK: Blackwell.
- Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, 12(Suppl.), preface, p. iii.
- Haberman, S. J. (1979). *Analysis of qualitative data: Vol. 2. New developments*. New York: Academic Press.
- Hacking, I. (1965). *The logic of statistical inference*. Cambridge, UK: Cambridge University Press.
- Hacking, I. (2001). *An introduction to probability and inductive logic*. Cambridge, UK: Cambridge University Press.
- Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear analysis of panel, trend and cohort data*. Newbury Park, CA: Sage.
- Hagenaars, J. A. (1993). *Loglinear models with latent variables*. Newbury Park, CA: Sage.
- Hagenaars, J. A. (1998). Categorical causal modeling: Latent class analysis and directed log-linear models with latent variables. *Sociological Methods and Research*, 26, 436–486.
- Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

- Hagle, T. M., & Mitchell, G. E., II. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 36, 762–784.
- Haight, B. K., Michel, Y., & Hendrix, S. (1998). Life review: Preventing despair in newly relocated nursing home residents: Short- and long-term effects. *International Journal of Aging and Human Development*, 47(2), 119–142.
- Hajek, J., & Sidák, Z. (1967). *Theory of rank tests*. New York: Academic Press.
- Hakim, C. (1982). *Secondary analysis of social research*. London: George Allen and Unwin.
- Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum.
- Hald, A. (1952). *Statistical theory with engineering applications*. New York: John Wiley.
- Hald, A. (1990). *A history of probability and statistics and their applications before 1750*. New York: John Wiley and Sons.
- Halfpenny, P. (1982). *Positivism and sociology: Explaining social life*. London: Allen & Unwin.
- Hall, E. T. (1974). *Handbook for proxemic research*. Washington, DC: Society for the Anthropology of Visual Communication.
- Hall, S. (1974). The television discourse—Encoding and decoding. In *Education and culture* (Vol. 35, pp. 8–14). Paris: United Nations Educational, Social, and Cultural Organization.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hamilton, D. (1980). Some contrasting assumptions about case study research and survey analysis. In H. Simons (Ed.), *Towards a science of the singular: Essays about case study in educational research and evaluation* (pp. 78–92). Norwich, UK: Centre for Applied Research in Education University of East Anglia.
- Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.
- Hamilton, L. C. (1990). *Modern data analysis: A first course in applied statistics*. Pacific Grove, CA: Brooks/Cole.
- Hamilton, L. C. (1992). *Regression with graphics: A second course in applied statistics*. Pacific Grove, CA: Brooks/Cole.
- Hamlyn, D. W. (1967). Empiricism. In P. Edwards (Ed.), *The encyclopedia of philosophy* (pp. 499–505). New York: Macmillan.
- Hammersley, M. (1989). *The dilemma of qualitative method: Herbert Blumer and the Chicago school of sociology*. London: Routledge and Kegan Paul.
- Hammersley, M. (1989). The problem of the concept: Herbert Blumer on the relationship between concepts and data. *Journal of Contemporary Ethnography*, 18, 133–160.
- Hammersley, M. (1990). *Reading ethnographic research*. London: Longman.
- Hammersley, M. (1992). *What's wrong with ethnography*. London: Routledge Kegan Paul.
- Hammersley, M. (1995). *The politics of social research*. London: Sage.
- Hammersley, M. (1995). Theory and evidence in qualitative research. *Quality and Quantity*, 29, 55–66.
- Hammersley, M. (1997). Qualitative data archiving: Some reflections on its prospects and problems. *Sociology*, 31(1), 131–142.
- Hammersley, M. (1999). Sociology, what's it for? A critique of the grand conception. *Sociological Research Online*, 4(3). Retrieved from <http://www.socresonline.org.uk/socresonline/4/3/hammersley.html>
- Hammersley, M. (2000). *Taking sides in social research*. London: Routledge.
- Hammersley, M. (2001). On “systematic” reviews of research literatures: A “narrative” reply to Evans and Benefield. *British Educational Research Journal*, 27(5), 543–554.
- Hammersley, M., & Atkinson, P. (1995). *Ethnography: Principles in practice*. London: Routledge.
- Hammond, M., Howarth, J., & Keat, R. (1991). *Understanding phenomenology*. Oxford, UK: Blackwell.
- Handcock, M. S., & Morris, M. (1999). *Relative distribution methods in the social sciences*. New York: Springer-Verlag.
- Haney, C., Banks, C., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology*, 1, 69–97.
- Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory*. New York: John Wiley & Sons.
- Hansen, P. R., & Johansen, S. (1998). *Workbook on cointegration*. Oxford, UK: Oxford University Press.
- Hantrais, L., & Mangen, S. (Eds.). (1996). *Cross-national research methods in the social sciences*. London: Pinter.
- Harary, F., Norman, D., & Cartwright, D. (1965). *Structural models for directed graphs*. New York: Free Press.
- Hardin, J. W., & Hilbe, J. M. (2003). *Generalized estimating equations*. London: Chapman & Hall.
- Harding, S. (1991). *Whose science? Whose knowledge?* Milton Keynes, UK: Open University Press.
- Harding, S. (1992). After the neutrality ideal: Science, politics, and “strong objectivity.” *Social Research*, 59, 568–587.
- Harding, S. (Ed.). (1987). *Feminism and methodology*. Milton Keynes, UK: Open University Press.
- Hardy, M. A. (1993). *Regression with dummy variables* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–093). Newbury Park, CA: Sage.
- Hardy, M. A., & Reynolds, J. (2003). Incorporating categorical information into regression models: the utility of dummy variables. In M. A. Hardy & A. Bryman (Eds.), *Handbook of data analysis*. London: Sage.
- Hare, R. D. (1966). Temporal gradient of fear arousal in psychopaths. *Journal of Abnormal and Social Psychology*, 70, 442–445.
- Harper, D. (2001). *Changing works: Visions of a lost agriculture*. Chicago: University of Chicago Press.
- Harper, D. (2002). Talking about pictures: A case for photo elicitation. *Visual Studies*, 17(1), 13–26.
- Harré, R. (1961). *Theories and things*. London: Sheed & Ward.
- Harré, R. (1986). *Varieties of realism*. Oxford, UK: Blackwell.

- Harré, R., & Secord, P. F. (1972). *The explanation of social behaviour*. Oxford, UK: Basil Blackwell.
- Harris, C. W. (1962). Some Rao-Guttman relationships. *Psychometrika*, 27, 247–263.
- Harris, J. A. (1913). On the calculation of intraclass and interclass coefficients of correlation from class moments when the number of possible combinations is large. *Biometrika*, 9, 446–472.
- Harris, M. (1976). History and significance of the emic/etic distinction. *Annual Review of Anthropology*, 5, 329–350.
- Hart, C. (1998). *Doing a literature review: Releasing the social science research imagination*. London: Sage.
- Hartigan, J. A., & Kleiner, B. (1981). Mosaics for contingency tables. In W. F. Eddy (Ed.), *Computer science and statistics: Proceedings of the 13th Symposium on the Interface* (pp. 268–273). New York: Springer-Verlag.
- Hartsock, N. C. M. (1998). *The feminist standpoint revisited and other essays*. Boulder, CO: Westview.
- Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis* (Sage University Papers on Quantitative Applications in the Social Sciences). Beverly Hills, CA: Sage.
- Harvey, A. (1990). *The econometric analysis of time series* (2nd ed.). Cambridge: MIT Press.
- Harvey, A. C. (1990). *Forecasting, structural time series models, and the Kalman filter*. New York: Cambridge University Press.
- Harvey, A. S. (1999). Guidelines. In W. E. Pentland, A. S. Harvey, M. P. Lawton, & M. A. McColl (Eds.), *Time use research in the social sciences* (pp. 19–45). New York: Kluwer.
- Harvey, D. (1973). *Social justice and the city*. Baltimore: Johns Hopkins University Press.
- Harvey, D. (1989). *The condition of postmodernity*. Oxford, UK: Basil Blackwell.
- Harvey, L. (1987). *Myths of the Chicago school of sociology*. Aldershot, England: Avebury.
- Hastie, T. J. (1992). Generalized additive models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S* (pp. 249–307). Pacific Grove, CA: Wadsworth and Brooks/Cole.
- Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman & Hall.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer-Verlag.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains. *Biometrika*, 57, 97–109.
- Hauser, R. M. (1978). A structural model of the mobility table. *Social Forces*, 56, 919–953.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46, 1251–1271.
- Hausman, J., & McFadden, D. (1984). Specification tests for the multinomial logit model. *Econometrica*, 52, 1219–1240.
- Hayano, D. (1979). Auto-ethnography: Paradigms, problems and prospects. *Human Organization*, 38(1), 99–104.
- Hayes, S. C., White, D., & Bissett, R. T. (1998). Protocol analysis and the “silent dog” method of analyzing the impact of self-generated rules. *Analysis of Verbal Behavior*, 15, 57–63.
- Hays, W. L. (1972). *Statistics for the social sciences*. New York: Holt.
- Hays, W. L. (1988). *Statistics*. New York: Holt, Rinehart and Winston.
- Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.
- Headland, T. N., Pike, K. L., & Harris, M. (1990). *Emics and etics: The insider/outsider debate*. Newbury Park, CA: Sage.
- Healey, J. F. (1995). *Statistics: A tool for social research* (3rd ed.). Belmont, CA: Wadsworth.
- Heaton, J. (1998). Secondary analysis of qualitative data [Online]. *Social Research Update*, 22. Retrieved April 28, 2002, from <http://www.soc.surrey.ac.uk/sru/SRU22.html>.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- Heckman, J. J., & Singer, B. (1982). Population heterogeneity in demographic models. In K. Land & A. Rogers (Eds.), *Multidimensional mathematical demography* (pp. 567–599). New York: Academic Press.
- Heckman, J. J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52(2), 271–320.
- Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, 9, 85–110.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hedstrom, P., & Swedberg, R. (1998). *Social mechanisms: An analytical approach to social theory*. Cambridge, UK: Cambridge University Press.
- Heider, K. G. (1976). *Ethnographic film*. Austin: University of Texas Press.
- Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks, CA: Sage.
- Heise, D. (1971). The semantic differential and attitude research. In G. F. Summers (Ed.), *Attitude measurement* (pp. 235–253). Chicago: Rand McNally.
- Heise, D. R. (1975). *Causal analysis*. New York: Wiley.
- Heise, D. R. (1987). Affect Control Theory: Concepts and model. *Journal of Mathematical Sociology*, 13, 1–33.
- Hektner, J. M., & Csikszentmihalyi, M. (2002). The experience sampling method: Measuring the context and the content of lives. In R. B. Bechtel & A. Churchman (Eds.), *Handbook of environmental psychology* (pp. 233–243). New York: John Wiley & Sons.
- Helland, I. S. (1990). PLS regression and statistical models. *Scandinavian Journal of Statistics*, 17, 97–114.
- Hempel, C. (1965). *Aspects of scientific explanation, and other essays in the philosophy of science*. New York: Free Press.

- Hempel, C. (1965). Confirmation, induction, and rational belief. In C. Hempel (Ed.), *Aspects of scientific explanation*. New York: Free Press.
- Hempel, C. G. (1966). *Philosophy of natural science*. Englewood Cliffs, NJ: Prentice Hall.
- Hendrick, T., Bickman, L., & Rog, D. J. (1993). *Applied research design: A practical guide*. Newbury Park, CA: Sage.
- Hendry, D. F. (1995). *Dynamic econometrics*. New York: Oxford University Press.
- Henkel, R. E. (1976). *Tests of significance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-004). Beverly Hills, CA: Sage.
- Henle, M., & Hubble, M. B. (1938). Egocentricity in adult conversation. *Journal of Social Psychology*, 9, 227-234.
- Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.
- Henry, G. T., & Gordon, C. S. (2001). Tracking issue attention: Specifying the dynamics of the public agenda. *Public Opinion Quarterly*, 65(2), 157-177.
- Hepburn, A. (2000). On the alleged incompatibility between feminism and relativism. *Feminism and Psychology*, 10, 91-106.
- Heritage, J. (1984). *Garfinkel and ethnomethodology*. Cambridge, UK: Polity.
- Herrnstein, R., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.
- Hess, I. (1985). *Sampling for social research surveys 1947-1980*. Ann Arbor, MI: Institute for Social Research.
- Hewitt, J. (1989). *Dilemmas of the American self*. Philadelphia: Temple University Press.
- Heyde, C. C., & Seneta, E. (Eds.). (2001). *Statisticians of the centuries*. New York: Springer.
- Hibbs, D. A. (1982). The dynamics of political support for American presidents among occupational and partisan groups. *American Journal of Political Science*, 26, 312-332.
- Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, 13, 141-154.
- Hill, A. B. (1953). Observation and experiment. *New England Journal of Medicine*, 248, 995-1001.
- Hill, M. M. (1993). *Archival strategies and techniques* (Qualitative Research Methods Series No. 31). Newbury Park, CA: Sage.
- Hill, M. S. (1992). *The panel study of income dynamics: A user's guide*. Newbury Park, CA: Sage.
- Hilsum, S., & Cane, B. (1971). *The teacher's day*. Windsor, UK: National Foundation for Educational Research.
- Himmelfarb, G. (1987). *The new history and the old*. Cambridge, MA: Harvard University Press.
- Hinchman, L. P., & Hinchman, S. K. (Eds.). (1997). *Memory, identity, community: The idea of narrative in the human sciences*. Albany: State University of New York Press.
- Hinds, P. S., Vogel, R. J., & Clarke-Steffen, L. (1997). The possibilities and pitfalls of doing a secondary analysis of a qualitative data set. *Qualitative Health Research*, 7, 408-424.
- Hine, C. (2000). *Virtual ethnography*. London: Sage.
- Hinton, P. R. (1996). *Statistics explained: A guide for social science students*. London: Routledge.
- Hirsch, M. W., & Smale, S. (1974). *Differential equations, dynamical systems, and linear algebra*. New York: Academic Press.
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.
- Hobcraft, J., Menken, J., & Preston, S. (1982). Age, period, and cohort effects in demography: A review. *Population Index*, 42, 4-43.
- Hobert, J. P., & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461-1473.
- Hobson-West, P., & Sapsford, R. (2001). *The Middlesbrough Town Centre Study: Final report*. Middlesbrough, UK: University of Teesside, School of Social Sciences.
- Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: John Wiley.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression. *Technometrics*, 12(1), 55-67, 69-82.
- Hoey, M. (2001). *Textual interaction*. London: Routledge.
- Hofmann, H. (2000). Exploring categorical data: Interactive mosaic plots. *Metrika*, 51(1), 11-26.
- Hogg, R. V., & Craig, A. T. (1978). *Introduction to mathematical statistics* (4th ed.). New York: Macmillan.
- Hogg, R. V., & Tanis, E. A. (1988). *Probability and statistical inference*. New York: Macmillan.
- Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical inference*. New York: Wiley.
- Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd ed.). New York: Wiley.
- Hollis, M. (1977). *Models of man*. Cambridge, UK: Cambridge University Press.
- Hollis, M., & Smith, S. (1990). *Explaining and understanding international relations*. New York: Oxford University Press.
- Hollway, W., & Jefferson, T. (2000). *Doing qualitative research differently: Free association, narrative and the interview method*. London: Sage.
- Holmes, D. (1976). Debriefing after psychological experiments: Effectiveness of post-experimental desensitizing. *American Psychologist*, 32, 868-875.
- Holmwood, J. (2001). Gender and critical realism: A critique of Sayer. *Sociology*, 35, 947-965.
- Holstein, J. A., & Gubrium, J. F. (1995). *The active interview*. Thousand Oaks, CA: Sage.
- Holstein, J. A., & Gubrium, J. F. (2000). *The self we live by: Narrative identity in a postmodern world*. New York: Oxford University Press.

- Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.
- Homan, R. (1992). *The ethics of social research*. London: Longman.
- Homans, G. C. (1964). Contemporary theory in sociology. In R. E. L. Faris (Ed.), *Handbook of modern sociology* (pp. 951–977). Chicago: Rand McNally.
- Homans, G. C. (1967). *The nature of social science*. New York: Harcourt, Brace, & World.
- Honneth, A. (1996). *The struggle for recognition: The moral grammar of social conflicts* (J. Anderson, Trans.). Cambridge: MIT Press.
- Hopkins, K. D. (1982). The unit of analysis: Group means versus individual observations. *American Educational Research Journal*, 19(1), 5–18.
- Horn, R. (1996). Negotiating research access to organizations. *Psychologist*, 9(12), 551–554.
- Horst, P. (1941). The role of the predictor variables which are independent of the criterion. *Social Science Research Council*, 48, 431–436.
- Horton, M., & Freire, P. (2000). *We make the road by walking* (Bell, B., Gaventa, J., & Peters, J., Eds.). Philadelphia: Temple University Press.
- Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician*, 55, 244–254.
- Horvitz, D. G., Shah, B. U., & Simmons, W. R. (1967). The unrelated question randomized response model. In E. D. Goldfield (Ed.), *Proceedings of the Social Statistics Section* (pp. 65–72). Washington, DC: American Statistical Association.
- Hoskins, J. (1998). *Biographical objects: How things tell the stories of people's lives*. London: Routledge.
- Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, 2, 211–228.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441, 498–520.
- Hougaard, P. (2000). *Analysis of multivariate survival data*. New York: Springer-Verlag.
- House, E. R., & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage.
- Hout, M. (1983). *Mobility tables*. Beverly Hills, CA: Sage.
- Howell, D. C. (1999). *Fundamental statistics for the behavioral sciences* (4th ed.). Belmont, CA: Duxbury.
- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Duxbury, UK: Thomson Learning.
- Hox, J. J. (2002). *Multilevel analysis, techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.
- Hoyle, R. H. (2000). Confirmatory factor analysis. In H. E. A. Tinsely & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 465–497). New York: Academic Press.
- Hsiao, C. (1986). *Analysis of panel data*. Cambridge, UK: Cambridge University Press.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1–55.
- Huber, P. (1981). *Robust statistics*. New York: Wiley.
- Huberman, A. M., & Miles, M. B. (1994). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 428–444). Thousand Oaks, CA: Sage.
- Huck, S. (2000). *Reading statistics and research* (3rd ed.). New York: Addison Wesley Longman.
- Huff, A. S. (2000). Changes in organizational knowledge production. *Academy of Management Review*, 25(2), 288–293.
- Hughes, J. A. (1990). *The philosophy of social research* (2nd ed.). Harlow, UK: Longman.
- Hughes, K., MacKintosh, A. M., Hastings, G., Wheeler, C., Watson, J., & Inglis, J. (1997). Young people, alcohol, and designer drinks: A quantitative and qualitative study. *British Medical Journal*, 314, 414–418.
- Hughes, R. (1998). Considering the vignette technique and its application to a study of drug injecting and HIV risk and safer behaviour. *Sociology of Health and Illness*, 20, 381–400.
- Hujala, E. (1998). Problems and challenges in cross-cultural research. *Acta Universitatis Ouluensis*, 35, 19–31.
- Hume, D. (1786). *Treatise on human nature*. Oxford, UK: Clarendon.
- Humphreys, L. (1975). *Tearoom trade: Impersonal sex in public places*. Chicago: Aldine.
- Hunt, J. C. (1989). *Psychoanalytic aspects of fieldwork* (University Paper Series on Qualitative Research Methods 18). Newbury Park, CA: Sage.
- Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, 86, 721–735.
- Husserl, E. (1970). *Logical investigations* (Vols. 1–2). London: Routledge Kegan Paul.
- Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. London: Sage.
- Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association*, 65, 1582–1589.
- Hyde, R. (2003). *The art of assembly language*. San Francisco: No Starch Press.
- Hyman, H. (1955). *Survey design and analysis*. New York: Free Press.
- Hyman, H. H. (1972). *Secondary analysis of sample surveys*. Glencoe, IL: Free Press.
- Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.
- ICPSR. (2002). *Guide to social science data preparation and archiving* [Online]. Available: http://www.ifdo.org/archiving_distribution/datprep_archiving_bfr.htm.
- IFDO. (n.d.). [Online]. Available: <http://www.ifdo.org/>.

- Imber, J. B. (Ed.). (2001). Symposium: Population politics. *Society*, 39, 3–53.
- Inkeles, A., & Sasaki, M. (Eds.). (1996). *Comparing nations and cultures*. Englewood Cliffs, NJ: Prentice Hall.
- International Centre for Diarrhoeal Disease Research, Bangladesh (ICDDR). (1992). *Demographic surveillance system—Matlab: Registration of demographic events—1985* (Scientific Report 68). Dhaka, Bangladesh: Author.
- Inter-University Consortium for Political and Social Research. (2002). *Guide to social science data preparation and archiving* [Online]. Supported by the Robert Wood Johnson Foundation. Available: www.ICPSR.umich.edu
- Iversen, G. R. (1973). Recovering individual data in the presence of group and individual effects. *American Journal of Sociology*, 79, 420–434.
- Iversen, G. R. (1984). *Bayesian statistical inference* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–043). Beverly Hills, CA: Sage.
- Iversen, G. R. (1991). *Contextual analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–081). Newbury Park, CA: Sage.
- Iversen, G. R., & Norpoth, H. (1987). *Analysis of variance* (2nd ed., Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–001). Newbury Park, CA: Sage.
- Iversen, G., & Gergen, M. (1997). *Statistics: The conceptual approach*. New York: Springer-Verlag.
- Iyengar, S., Peters, M. E., & Kinder, D. R. (1982). Experimental demonstrations of the “not-so-minimal” consequences of television news programs. *The American Journal of Political Science*, 4, 848–858.
- Jaccard, J. (1998). *Interaction effects in factorial analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–118). Thousand Oaks, CA: Sage.
- Jaccard, J., & Turrissi, R. (2003). *Interaction effects in multiple regression* (2nd ed., Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–72). Thousand Oaks, CA: Sage.
- Jaccard, J., Turrissi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. Newbury Park, CA: Sage.
- Jackman, R. W. (1973). On the relationship of economic development to political performance. *American Journal of Political Science*, 17, 611–621.
- Jackson, D. N. (2002). The constructs in people’s heads. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 3–18). Mahwah, NJ: Lawrence Erlbaum.
- Jackson, J. E. (1991). *A user’s guide to principal component analysis*. New York: John Wiley.
- Jackson, S., & Brashers, D. E. (1994). *Random Factors in ANOVA* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–098). Thousand Oaks, CA: Sage.
- Jacobs, H. A. (1861). *Incidents in the life of a slave girl: Written by herself*. Boston.
- Jacoby, L. L., Debnar, J. A., & Hay, J. F. (2001). Proactive interference, accessibility bias, and process dissociations: Valid subject reports of memory. *Journal of Experimental Psychology: Memory, Learning, and Cognition*, 27, 686–700.
- Jacoby, W. G. (1991). *Data theory and dimensional analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–078). Newbury Park, CA: Sage.
- Jacoby, W. G. (1997). *Statistical graphics for univariate and bivariate data*. Thousand Oaks, CA: Sage.
- Jaeger, R. M. (1993). *Statistics: A spectator sport* (2nd ed.). Newbury Park, CA: Sage.
- Jahoda, G. (1977). In pursuit of the emic-etic distinction: Can we ever capture it? In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural research* (pp. 55–63). Amsterdam: Swets & Zeitlinger.
- James, J. B., & Sørensen, A. (2000, December). Archiving longitudinal data for future research: Why qualitative data add to a study’s usefulness. *Forum Qualitative Sozialforschung* [Forum: Qualitative Social Research] [Online journal], 1(3). Available: <http://qualitative-research.net/fqs/fqs-eng.htm>
- James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69, 85–98.
- James, W. (1952). *The varieties of religious experience: A study in human nature*. London: Longmans, Green & Co. (Original work published 1902)
- Jameson, F. (1983). Postmodernism and consumer society. In H. Foster (Ed.), *Postmodern culture*. London: Pluto.
- Jasso, G. (1988). Principles of theoretical analysis. *Sociological Theory*, 6, 1–20.
- Jasso, G. (1990). Methods for the theoretical and empirical analysis of comparison processes. *Sociological Methodology*, 20, 369–419.
- Jasso, G. (2001). Formal theory. In J. H. Turner (Ed.), *Handbook of sociological theory* (pp. 37–68). New York: Kluwer Academic/Plenum.
- Jasso, G. (2001). Rule-finding about rule-making: Comparison processes and the making of norms. In M. Hechter & K.-D. Opp (Eds.), *Social norms* (pp. 348–393). New York: Russell Sage.
- Jasso, G. (2001). Comparison theory. In J. H. Turner (Ed.), *Handbook of sociological theory* (pp. 669–698). New York: Kluwer Academic/Plenum.
- Jasso, G. (2001). Studying status: An integrated framework. *American Sociological Review*, 66, 96–124.
- Jasso, G. (2002). Seven secrets for doing theory. In J. Berger & M. Zelditch (Eds.), *New directions in contemporary sociological theory* (pp. 317–342). Boulder, CO: Rowan & Littlefield.
- Jasso, G. (in press). The tripartite structure of social science analysis. *Sociological Theory*.
- Jasso, G., & Rossi, P. H. (1977). Distributive justice and earned income. *American Sociological Review*, 42, 639–651.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Clarendon.

- Jenkins, G. M., & Watts, D. G. (1968). *Spectral analysis and its applications*. San Francisco: Holden-Day.
- Jenkins, R. (1984). Bringing it all back home: An anthropologist in Belfast. In C. Bell & H. Roberts (Eds.), *Social researching: Politics, problems, practice* (pp. 147–164). London: Routledge and Kegan Paul.
- JMP Introductory Guide. (2000). Cary, NC: SAS Institute.
- Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12, 231–254.
- Johansen, S. (1991). Estimation and hypotheses testing of cointegrating vectors in Gaussian vector autoregressive models. *Econometrica*, 59, 1551–1580.
- Johansen, S. (1995). *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford, UK: Oxford University Press.
- John, K. E. (1989). The polls—A report. *Public Opinion Quarterly*, 53, 590–605.
- John, N. R., & Draper, J. A. (1980). An alternative family of transformations. *Applied Statistics*, 29(2), 190–197.
- Johnson, C. (1982). Risks in the publication of fieldwork. In J. E. Sieber (Ed.), *The ethics of social research: Fieldwork, regulation and publication* (pp. 71–92). New York: Springer-Verlag.
- Johnson, C. (1994). Gender, legitimate authority, and leader-subordinate conversations. *American Sociological Review*, 59, 122–135.
- Johnson, J. C. (1990). *Selecting ethnographic informants*. Newbury Park, CA: Sage.
- Johnson, J. C., & Weller, S. C. (2002). Elicitation techniques for interviewing. In J. F. Gubrium & J. A. Holstein (Eds.), *The handbook of interview research*. Thousand Oaks, CA: Sage.
- Johnson, N. L., & Kotz, S. (1969). *Discrete distributions*. Boston: Houghton-Mifflin.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions, Vol. 1*. New York: Wiley.
- Johnson, R. A., & Bhattacharyya, G. K. (2001). *Statistics: Principles and methods* (4th ed.). New York: Wiley.
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Upper Saddle River, NJ: Prentice Hall.
- Johnson, T., Dandeker, C., & Ashworth, C. (1984). *The structure of social theory*. London: Macmillan.
- Johnston, J., & Dinardo, J. (1997). *Econometric methods* (4th ed.). New York: McGraw-Hill.
- Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards*. Thousand Oaks, CA: Sage.
- Joliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.
- Jones, J. M. G., & Hunter, D. (1995). Consensus methods for medical and health services research. *British Medical Journal*, 311, 376–380.
- Jones, K. (1997). Multilevel approaches to modelling contextuality: From nuisance to substance in the analysis of voting behaviour. In G. P. Westert & R. N. Verhoeff (Eds.), *Places and people: Multilevel modelling in geographical research* (Nederlandse Geografische Studies 227). Utrecht, The Netherlands: The Royal Dutch Geographical Society and Faculty of Geographical Sciences, Utrecht University.
- Jones, M. B. (1959). *Simplex theory* (U.S. Naval School of Aviation Medicine Monograph Series No. 3). Pensacola, FL: U.S. Naval School of Aviation Medicine.
- Jones, M. O. (1996). *Studying organizational symbolism: What, how, why?* Thousand Oaks, CA: Sage.
- Jones, S. R. (1992). Was there a Hawthorne effect? *American Journal of Sociology*, 98, 451–468.
- Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32, 443–482.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183–202.
- Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York: Academic Press.
- Jöreskog, K. G. (1979). Statistical models and methods for analysis of longitudinal data. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt.
- Josselson, R., & Lieblich, A. (Eds.). (1993–1999). *The narrative study of lives* (Vols. 1–6). Newbury Park, CA: Sage.
- Josselsyn, R. (Ed.). (1996). *Ethics and process in the narrative study of lives*. Thousand Oaks, CA: Sage.
- Joynson, R. B. (1989). *The Burt affair*. London: Routledge.
- Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1). Boston: McGraw-Hill.
- Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). Fort Worth, TX: Harcourt Brace Jovanovich.
- Judge, G. G., Griffiths, W. E., Hill, R. C., & Lee, T.-C. (1980). *The theory and practice of econometrics* (Wiley Series in Probability and Mathematical Statistics). New York: John Wiley.
- Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., & Lee, T.-C. (1985). *The theory and practice of econometrics* (2nd ed.). New York: John Wiley.
- Judge, G. G., Hill, R. C., Griffiths, W. E., Lütkepohl, H., & Lee, T. (1988). *Introduction to the theory and practice of econometrics* (2nd ed.). New York: John Wiley.
- Julian, D. A. (1997). The utilization of the logic model as a system level planning and evaluation device. *Evaluation and Program Planning*, 20(3), 251–257.
- Juster, F. T., & Stafford, F. P. (1985). *Time, goods and well-being*. Ann Arbor, MI: Institute for Social Research.
- Kalichman, S. C. (1999). *Mandated reporting of suspected child abuse: Ethics, law and policy*. Washington, DC: American Psychological Association.

- Kalton, G. (1983). *Introduction to survey sampling* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-035). Beverly Hills, CA: Sage.
- Kan, M. (2002). Reinterpreting the Multifactor Leadership Questionnaire. In K. W. Parry & J. R. Meindl (Eds.), *Grounding leadership theory and research: Issues, perspectives and methods* (pp. 159-173). Greenwich, CT: Information Age Publishing.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31-41.
- Kant, I. (1781). *Kritik der reinen Vernunft* [Critique of pure reason]. Hamburg: Felix Meiner.
- Kanuha, V. K. (2000). "Being" native versus "going native": Conducting social work research as an insider. *Social Work*, 45(5), 439-447.
- Kaplan, A. (1943). Content analysis and the theory of signs. *Philosophy of Science*, 10, 230-247.
- Kaplan, A. (1964). *The conduct of inquiry*. New York: Chandler.
- Kaplan, D. (1990). Evaluation and modification of covariance structure models: A review and recommendation. *Multivariate Behavioral Research*, 25, 137-155.
- Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-481.
- Kashy, D. A., & Kenny, D. A. (2000). The analysis of data from dyads and groups. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 451-477). New York: Cambridge University Press.
- Kasprzyk, D., Duncan, G. J., Kalton, G., & Singh, M. P. (1989). *Panel surveys*. New York: John Wiley.
- Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2), 119-127.
- Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343-1370.
- Kazdin, A. E. (2003). *Research design in clinical psychology, 4th Edition*. Boston: Allyn & Bacon.
- Keat, R., & Urry, J. (1975). *Social theory as science*. London: Routledge Kegan Paul.
- Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125-148.
- Keleman, M., & Bansal, P. (2002). The conventions of management research and their relevance to management practice. *British Journal of Management*, 13, 97-108.
- Kelle, U. (Ed.). (1995). *Computer-aided qualitative data analysis: Theory, methods and practice*. London: Sage.
- Kelly, G. (1955). *The psychology of personal constructs*. New York: Norton.
- Kendall, G., & Wickham, G. (1999). *Using Foucault's methods*. London: Sage.
- Kendall, M. G. (1962). *Rank correlation methods* (3rd ed.). London: Griffin.
- Kendall, M. G., & Buckland, W. R. (1971). *A dictionary of statistical terms* (3rd ed.). Edinburgh: Oliver & Boyd.
- Kendall, M., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). New York: Oxford University Press.
- Kendall, M., & Stuart, A. (1969). *The advanced theory of statistics: Vol. 2. Inference and relationship*. London: Griffin.
- Kendall, P. L., & Lazarsfeld, P. F. (1950). Problems of survey analysis. In R. K. Merton & P. F. Lazarsfeld (Eds.), *Continuities in social research: Studies in the scope and method of "the American soldier"* (pp. 160-176). Glencoe, IL: Free Press.
- Kennedy, P. (1992). *A guide to econometrics* (3rd ed.). Cambridge: MIT Press.
- Kennedy, P. (1998). *A guide to econometrics* (4th ed.). Cambridge: MIT Press.
- Kennedy, P. E. (2002). More on Venn diagrams for regression. *Journal of Statistics Education*, 10(1) [Online]. Retrieved from <http://www.amstat.org/publications/jse/v10n1/kennedy.html>
- Kenny, D. A. (1979). *Correlation and causality*. New York: John Wiley.
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford.
- Kenny, D. A., & La Voie, L. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 18, pp. 142-182). Orlando, FL: Academic Press.
- Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Keren, G., & Lewis, C. (Eds.). (1993). *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Lawrence Erlbaum.
- Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research*. Ft. Worth, TX: Harcourt.
- Kerr, A. W., Hall, H. K., & Kozub, A. (2002). Descriptive statistics. Chapter 2 in *Doing statistics with SPSS*. London: Sage.
- Keselman, H. J., Lix, L. M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods*, 3, 123-141.
- Kessler, R., & Greenberg, D. (1981). *Linear panel analysis*. New York: Academic Press.
- Keyfitz, N. (1985). *Applied mathematical demography* (2nd ed.). New York: Wiley.
- Keynes, J. M. (1936). *The general theory of employment, interest, and money*. New York: Harcourt Brace Jovanovich.
- Khurana, B. (1995). *The older spouse caregiver: Paradox and pain of Alzheimer's disease*. Unpublished doctoral dissertation, Center for Psychological Studies, Albany, CA.
- Kiel, L. D., & Elliott, E. (1996). *Chaos theory in the social sciences*. Ann Arbor: University of Michigan Press.
- Kim, J. O., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues*. Beverly Hills, CA: Sage.

- Kinal, T., & Lahiri, K. (1983). Specification error analysis with stochastic regressors. *Econometrica*, 54, 1209–1220.
- Kincaid, H. (1994). Defending laws in the social sciences. In M. Martin & L. McIntyre (Eds.), *Readings in the philosophy of social science* (pp. 111–130). Cambridge: MIT Press.
- Kincaid, H. (1996). *Philosophical foundations of the social sciences: Analyzing controversies in social research*. Cambridge, UK: Cambridge University Press.
- King, G. (1988). Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential Poisson regression model. *American Journal of Political Science*, 32, 838–863.
- King, G. (1989). *Unifying political methodology: The likelihood theory of statistical inference*. New York: Cambridge University Press.
- King, G. (1990). Stochastic variation: A comment on Lewis-Beck and Skalaban's "The R-Squared." *Political Analysis*, 2, 185–200.
- King, G. (1997). *A solution to the ecological inference problem*. Princeton, NJ: Princeton University Press.
- King, G. (1998). *Unifying political methodology: The likelihood theory of statistical inference*. Ann Arbor: University of Michigan Press.
- King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9, 137–163.
- King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.
- King, J. A. (1998). Making sense of participatory evaluation practice. *New Directions for Evaluation*, 80, 57–67.
- Kirk, J., & Miller, M. L. (1986). *Reliability and validity in qualitative research*. Beverly Hills, CA: Sage.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
- Kirk, R. E. (1999). *Statistics: An introduction* (4th ed.). Orlando, FL: Harcourt Brace.
- Kirkman, B. L., Rosen, B., Gibson, C. B., Tesluk, P. E., & McPherson, S. O. (2003). Five challenges to virtual team success: Lessons from Sabre, Inc. *Academy of Management Executive*, 16(3), 67–79.
- Kish, L. (1949). A procedure for objective respondent selection within a household. *Journal of the American Sociological Association*, 44, 380–387.
- Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92–115.
- Kish, L. (1965). *Survey sampling*. New York: Wiley.
- Kish, L. (1987). *Statistical design for research*. New York: Wiley.
- Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11(1), 55–77.
- Kitagawa, E. (1955). Components of a difference between two rates. *Journal of the American Statistical Association*, 50, 1168–1194.
- Kitcher, P., & Salmon, W. C. (Eds.). (1989). *Scientific explanation* (Minnesota Studies in the Philosophy of Science, Vol. 13). Minneapolis: University of Minnesota Press.
- Kivlahan, D. R., Marlatt, G. A., Fromme, K., Coppel, D. B., & Williams, E. (1990). Secondary prevention with college drinkers: Evaluation of an alcohol skills training program. *Journal of Consulting and Clinical Psychology*, 58(6), 805–810.
- Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. Hoboken, NJ: Wiley.
- Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic research: Principles and quantitative methods*. New York: Van Nostrand Reinhold.
- Kleinman, A. L., Eisenberg, L., & Good, B. J. (1978). Culture, illness and care: Clinical lessons from anthropologic and cross-cultural research. *Annals of Internal Medicine*, 88, 251–258.
- Kleinman, S., & Copp, M. A. (1993). *Emotions and fieldwork*. Thousand Oaks, CA: Sage.
- Kline, P. (1994). *An easy guide to factor analysis*. Thousand Oaks, CA: Sage.
- Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.
- Klir, G. J., & Folger, T. A. (1988). *Fuzzy sets, uncertainty, and information*. Englewood Cliffs, NJ: Prentice Hall.
- Kmenta, J. (1997). *Elements of econometrics* (2nd ed.). Ann Arbor: University of Michigan Press.
- Knoke, D., & Burke, P. J. (1980). *Log-linear models*. Beverly Hills, CA: Sage.
- Knoke, D., Bohrnstedt, G. W., & Mee, A. P. (2002). *Statistics for social data analysis* (4th ed.). Itasca, IL: Peacock.
- Knuth, D. E. (1997). *Fundamental algorithms: The art of computer programming* (Vol. 1, 3rd ed.). Reading, MA: Addison-Wesley.
- Knuth, D. E. (1997). *The art of computer programming, volume 2: Seminumerical algorithms*. Reading, MA: Addison-Wesley.
- Koch, W., Schulz, E. M., Wright, R., Smith, R. M., Lang, S. et al. (1996). What is a ratio scale? *Rasch Measurement Transactions*, 9, 457.
- Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (Eds.). (2000). *To err is human: Building a safer health system*. Washington, DC: National Academy Press.
- Kondo, D. K. (1990). *Crafting selves: Power, gender and discourses of identity in a Japanese workplace*. Chicago: University of Chicago Press.
- Korczynski, M. (2000). The political economy of trust. *Journal of Management Studies*, 37(1), 1–22.
- Kornberg, A. (1997). *Basic research, the lifeline of medicine*. Retrieved from <http://www.nobel.se/medicine/articles/research/>
- Körner, S. (1955). *Kant*. Harmondsworth, UK: Penguin.
- Kotz, S., & Johnson, N. (Eds.). (1983). *Encyclopedia of statistical sciences*. New York: Wiley.

- Kotz, S., Johnson, N. L., & Read, C. B. (1982). Censoring. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 1, p. 396). New York: Wiley.
- Koutsoyiannis, A. (1978). *Theory of econometrics: An introductory exposition of econometric methods* (2nd ed.). London: Macmillan.
- Kozlowski, S. W., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology*, 77, 161–167.
- Kracauer, S. (1952–1953). The challenge of qualitative content analysis. *Public Opinion Quarterly*, 16, 631–642.
- Krantz, J. H., & Dalal, R. (2000). Validity of Web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–60). New York: Academic Press.
- Krieger, N. (1994). Epidemiology and the web of causation: Has anyone seen the spider? *Social Science & Medicine*, 39(7), 887–903.
- Krieger, S. (1983). *The mirror's dance: Identity in a women's community*. Philadelphia: Temple University Press.
- Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage.
- Kristeva, J. (1991). *Strangers to ourselves*. New York: Columbia University Press.
- Kroenke, D. M. (2001). *Database processing: Fundamentals, design and implementation*. Upper Saddle River, NJ: Prentice Hall.
- Krosnick, J. A. (1999). Survey methodology. *Annual Review of Psychology*, 50, 537–567.
- Krueger, R. A. (1994). *Focus groups: A practical guide for applied research* (2nd ed.). Thousand Oaks, CA: Sage.
- Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research* (3rd ed.). Thousand Oaks, CA: Sage.
- Kruskal, J. B. (1964). Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–27, 115–129.
- Kruskal, J. B. (1968). Transformation of data. In D. L. Sills (Ed.), *International encyclopedia of the social sciences* (Vol. 15, pp. 182–192). New York: Macmillan.
- Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53, 814–861.
- Kuhn, M., & McPartland, T. S. (1954). An empirical investigation of self attitudes. *American Sociological Review*, 19, 68–76.
- Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press. (Original work published 1962)
- Kuran, T. (1995). The inevitability of future revolutionary surprises. *American Journal of Sociology*, 100(6), 1528–1551.
- Kuusela, H., Spence, M. T., & Kanto, A. J. (1998). Expertise effects on prechoice decision processes and final outcomes: A protocol analysis. *European Journal of Marketing*, 32, 5–6, 559.
- Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage.
- Kvale, S. (1999). The psychoanalytic interview as qualitative research. *Qualitative Inquiry*, 5(1), 87–113.
- Labov, W. (1982). Speech actions and reactions in personal narrative. In D. Tannen (Ed.), *Analyzing discourse: Text and talk*. Washington, DC: Georgetown University Press.
- Lachenbruch, P. A. (1975). *Discriminant analysis*. New York: Hafner.
- LaFreniere, P., & Charlesworth, W. R. (1983). Dominance, attention, and affiliation in a preschool group: A nine-month longitudinal study. *Ethology and Sociobiology*, 4(2), 55–67.
- Lagopoulos, A. Ph., & Boklund-Lagopolou, K. (1992). *Meaning and geography: The social conception of the region in northern Greece*. Berlin: Mouton de Gruyter.
- Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixture distribution. *Journal of the American Statistical Association*, 73, 805–811.
- Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers. Vol. 1*. Cambridge, UK: Cambridge University Press.
- Lakatos, I., & Musgrave, A. (Eds.). (1970). *Criticism and the growth of knowledge*. Cambridge, UK: Cambridge University Press.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
- Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics*, 34, 1–14.
- Langeheine, R., & van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods & Research*, 18, 416–441.
- Langeheine, R., & Van de Pol, F. (1994). Discrete-time mixed Markov latent class models. In A. Dale & R. B. Davies (Eds.), *Analyzing social and political change: A casebook of methods* (pp. 171–197). London: Sage.
- Langellier, K. M. (2001). Personal narrative. In M. Jolly (Ed.), *Encyclopedia of life writing: Autobiographical and biographical forms* (Vol. 2). London: Fitzroy Dearborn.
- Langellier, K. M., & Peterson, E. E. (2003). *Performing narrative: The communicative practice of storytelling*. Philadelphia: Temple University Press.
- Lapadat, J. C., & Lindsay, A. C. (1999). Transcription in research and practice: From standardization of technique to interpretive positionings. *Qualitative Inquiry*, 5(1), 64–86.
- Larsen, R. J., & Marx, M. L. (1990). *Statistics*. Englewood Cliffs, NJ: Prentice Hall.
- Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon symposium* (pp. 112–136). New York: John Wiley.
- Latham, G. P., Skarlicki, D., Irvine, D., & Siegel, J. P. (1993). The increasing importance of performance appraisals to employee effectiveness in organizational settings in North America. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational*

- psychology 1993 (Vol. 8, pp. 87–132). Chichester, UK: Wiley.
- Lather, P. (1993). Fertile obsession: Validity after post-structuralism. *Sociological Quarterly*, 35, 673–694.
- Lather, P. (1999). To be of use: The work of reviewing. *Review of Educational Research*, 69(1), 2–7.
- Laudan, L. (1977). *Progress and its problems: Toward a theory of scientific growth*. Berkeley: University of California Press.
- Lauder, M. (2003). Covert participant observation of a deviant community: Justifying the use of deception. *Journal of Contemporary Religion*, 18(2), 185–196.
- Laufer, R. S., & Wolfe, M. (1977). Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of Social Issues*, 33, 44–87.
- Laurie, H., Smith, R., & Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, 15(2), 269–282.
- Lave, C., & March, J. G. (1978). *An introduction to models in the social sciences*. New York: Harper & Row.
- Lavrakas, P. J. (1987). *Telephone survey methods*. Newbury Park, CA: Sage.
- Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.
- Lavrakas, P. J. (1998). Methods for sampling and interviewing in telephone surveys. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 429–472). Thousand Oaks, CA: Sage.
- Lawlis, G. F., & Lu, E. (1972). Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin*, 78, 17–20.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563–575.
- Lax, D. A. (1985). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, 80, 736–741.
- Layder, D. (1993). *New strategies in social research*. Cambridge, UK: Polity.
- Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis & the interpretation and mathematical foundation of latent structure analysis. In S. A. Stouffer (Ed.), *Measurement and prediction* (pp. 362–472). Princeton, NJ: Princeton University Press.
- Lazarsfeld, P. F. (1955). Interpretation of statistical relations as a research operation. In P. F. Lazarsfeld & M. Rosenberg (Eds.), *The language of social research* (pp. 111–125). Glencoe, IL: Free Press.
- Lazarsfeld, P. F. (1958). Evidence and inference in social research. *Daedalus*, 87, 120–121.
- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.
- Leach, C., Freshwater, K., Aldridge, J., & Sunderland, J. (2001). Analysis of repertory grids in clinical practice. *British Journal of Clinical Psychology*, 40, 225–248.
- Leamer, E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York: Wiley.
- Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York: John Wiley.
- Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, 73(1), 31–43.
- Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive statistical analysis*. New York: Wiley.
- Lebreton, J.-D., Burnham, K. P., Clobert, J., & Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies. *Ecological Monographs*, 62, 67–118.
- Lee, A. M. (1978). *Sociology for whom*. New York: Oxford University Press.
- Lee, P. M. (1997). *Bayesian statistics: An introduction* (2nd ed.). London: Arnold.
- Lee, R. (2000). *Unobtrusive methods in social research*. Buckingham, UK: Open University Press.
- Lee, R. B. (1993). *The Dobe !Kung*. New York: Holt, Rinehart and Winston.
- Lee, R. M. (1993). *Doing research on sensitive topics*. London: Sage.
- Lee, R. M. (1995). *Dangerous fieldwork*. London: Sage.
- Lehtonen, M. (2000). *The cultural analysis of texts*. London: Sage.
- Leisering, L., & Leibfried, S. (1999). *Time and poverty in Western welfare states*. New York: Cambridge University Press.
- Lemaitre, G. (1992). *Dealing with the seam problem* (Survey of Labour and Income Dynamics Research Papers 92–05). Ottawa: Statistics Canada.
- Lenski, G. (1966). *Power and privilege*. New York: McGraw-Hill.
- Lepkowski, J. M., & Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 259–272). New York: John Wiley.
- Leslie, P. H. (1945). On the use of matrices in certain population mathematics. *Biometrika*, 33, 183–212.
- Lett, J. (1987). The importance of the emic/etic distinction. In *The human enterprise: A critical introduction to anthropological theory*. Boulder, CO: Westview.
- Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278–292). Stanford, CA: Stanford University Press.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10, 707–710. (Original work published 1965)
- Levin, A., Liukkonen, J., & Levine, D. W. (1996). Equivalent inference using transformations. *Communications in Statistics, Theory and Methods*, 25(5), 1059–1072.
- Levin, I. P. (1999). *Relating statistics and experimental design: An introduction* (Sage University Paper Series on

- Quantitative Applications in the Social Sciences, 07–125). Thousand Oaks, CA: Sage.
- Levin, J. R., & Subkoviak, M. J. (1977). Planning an experiment in the company of measurement error. *Applied Psychological Measurement*, 1(3), 331–338.
- Levinas, E. (1998). *On thinking-of-the-Other: Entre nous*. New York: Columbia University Press.
- Levine, E. K. (1982). Old people are not all alike: Social class, ethnicity/race, and sex are bases for important differences. In J. E. Sieber (Ed.), *The ethics of social research: Surveys and experiments* (pp. 127–144). New York: Springer-Verlag.
- Lévi-Strauss, C. (1963). *Structural anthropology*. New York: Basic Books.
- Lévi-Strauss, C. (1969). *The elementary structures of kinship*. London: Eyre and Spottiswoode.
- Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications*. New York: Wiley.
- Lewin, K. (1943). Defining the “field at a given time.” *Psychological Review*, 50, 292–310.
- Lewin, K. (1959). *A dynamic theory of personality: Selected papers* (D. K. Adams & K. E. Zener, Trans.). New York: McGraw-Hill. (Original work published 1935)
- Lewin, K. (1997). Experiments in social space. In G. W. Lewin (Ed.), *Resolving social conflicts: Field theory in social science* (pp. 59–67). Washington, DC: American Psychological Association. (Originally published in 1939)
- Lewis, D. K. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.
- Lewis-Beck, M. S. (1980). *Applied regression: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series 07–022). Beverly Hills, CA: Sage.
- Lewis-Beck, M. S. (1995). *Data analysis: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–103). Thousand Oaks, CA: Sage.
- Lewis-Beck, M. S. (Ed.). (1993). *Regression analysis*. London: Sage/Toppan.
- Lewis-Beck, M. S., & Skalaban, A. (1990). The R-squared: Some straight talk. *Political Analysis*, 2, 153–171.
- Li, H., Rosenthal, R., & Rubin, D. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, 1(1), 98–107.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.
- Liao, T. F. (1989). A flexible approach for the decomposition of rate differences. *Demography*, 26, 717–726.
- Liao, T. F. (1994). *Interpreting probability models: Logit, probit, and other generalized linear models*. Thousand Oaks, CA: Sage.
- Liao, T. F. (2001). How responsive is U.S. population growth to immigration? A situational sensitivity analysis. *Mathematical Population Studies*, 9, 217–229.
- Liao, T. F. (2002). *Statistical group comparison*. New York: Wiley.
- Lichfield, N. (1996). *Community impact evaluation*. London: University College Press.
- Lieberson, S. (1987). *Making it count: The improvement of social research and theory*. Berkeley: University of California Press.
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 44–53.
- Lincoln, Y. S., & Guba, E. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.
- Lincoln, Y. S. & Guba, E. (2000). Paradigmatic controversies, contradictions, and emerging confluences. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 163–188). Thousand Oaks, CA: Sage.
- Lind, J. (1753). *A treatise of the scurvy: Of three parts containing an inquiry into the nature, causes and cure of that disease*. Edinburgh, UK: Sands, Murray and Cochran.
- Lindahl, L. (2002). *Do birth order and family size matter for intergenerational income mobility? Evidence from Sweden* (Working Paper No. 5). Stockholm: Swedish Institute for Social Research.
- Lindell, M. K., & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of the CVI, T, $r_{WG(J)}$, and $r_{WG(J)}^*$ indexes. *Journal of Applied Psychology*, 84, 640–647.
- Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement*, 23, 127–135.
- Lindesmith, A. (1968). *Addiction and opiates*. Chicago: Aldine.
- Lindley, D. V. (2001). Thomas Bayes. In C. C. Heyde & E. Seneta (Eds.), *Statisticians of the centuries* (pp. 68–71). New York: Springer-Verlag.
- Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society*, 34, 1–41.
- Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.
- Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86, 96–107.
- Lingis, A. (1994). *Abuses*. Berkeley: The University of California Press.
- Lingis, A. (1994). *The community of those who have nothing in common*. Bloomington: Indiana University Press.
- Lingis, A. (1998). *The imperative*. Bloomington: Indiana University Press.
- Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). New York: Macmillan.
- Linstone, H. A. (1978). The Delphi technique In R. B. Fowles (Ed.), *Handbook of futures research*. Westport, CT: Greenwood.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute, Inc.

- Little, D. (1991). *Varieties of social explanation: An introduction to the philosophy of social science*. Boulder, CO: Westview.
- Little, D. (1998). *Microfoundations, method and causation: On the philosophy of the social sciences*. New Brunswick, NJ: Transaction Publishers.
- Little, R. J. A. (1988). Missing data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287–301.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley.
- Litwin, M. S. (1995). *How to measure survey reliability and validity*. Thousand Oaks, CA: Sage.
- Livingston, G. (1999). Beyond watching over established ways: A review as recasting the literature, recasting the lived. *Review of Educational Research*, 69(1), 9–19.
- Loader, C. (1999). *Local regression and likelihood*. New York: Springer.
- Locke, K. D. (2000). *Using grounded theory in management research*. London: Sage.
- Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(Suppl. 9), 635–694.
- Loewenthal, L. (1989). Sociology of literature in retrospect. In P. Desan et al. (Eds.), *Literature and social practice*. Chicago and London: University of Chicago Press.
- Lofland, J., & Lofland, L. H. (1995). *Analyzing social settings: A guide to qualitative observation and analysis*. Belmont, CA: Wadsworth.
- Loman, L. A., & Larkin, W. E. (1976). Rejection of the mentally ill: An experiment in labeling. *Sociological Quarterly*, 17, 555–560.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Long, J. S., & Cheng, S. (2003). Regression models for categorical outcomes. In M. A. Hardy & A. Bryman (Eds.), *Handbook of data analysis*. London: Sage.
- Long, J. S., & Ervin, L. H. (2000) Using heteroskedasticity consistent standard errors in the linear regression model. *American Statistician*, 54, 217–224.
- Longford, N. T. (1993). *Random coefficient models*. New York: Oxford University Press.
- Looney, C. G. (1997). *Pattern recognition using neural networks*. Oxford, UK: Oxford University Press.
- Lord, F. M. (1956). Sampling error due to choice of split in split-half reliability coefficients. *Journal of Experimental Education*, 24, 245–249.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lorencz, B. (1991). Becoming ordinary: Leaving the psychiatric hospital. In J. M. Morse & J. Johnson (Eds.), *The illness experience: Dimensions of suffering* (pp. 140–200). Newbury Park, CA: Sage.
- Lorenz, E. N. (1963). Deterministic non-periodic flow. *Journal of Atmospheric Science*, 20, 130–141.
- Lorenz, K. (1981). *The foundations of ethology*. New York: Simon and Schuster.
- Lotka, A. J. (1998). *Analytical theory of biological populations: Part II: Demographic analysis with particular application to human populations* (D. P. Smith & H. Rossert, Trans.). New York: Plenum. (Original work published 1939)
- Luger, G. F. (2002). *Artificial intelligence: Structures and strategies for complex problem solving* (4th ed.). Reading, MA: Addison-Wesley.
- Lukes, S. (1994). Methodological individualism reconsidered. In M. Martin & L. McIntyre (Eds.), *Readings in the philosophy of social science* (pp. 451–459). Cambridge: MIT Press.
- Lundberg, G. (1939). *Foundations of sociology*. New York: Macmillan.
- Lupia, A., McCubbins, M. D., & Popkin, S. L. (Eds.). (1998). *Elements of reason*. Cambridge, UK: Cambridge University Press.
- Luttrell, W. (2003). *Pregnant bodies, fertile minds: Gender, race, and the schooling of pregnant teens*. New York: Routledge.
- Lykken, D. T. (1957). A study of anxiety in the sociopathic personality. *Journal of Abnormal and Social Psychology*, 55, 6–10.
- Lykken, D. T. (1995). *The antisocial personalities*. Mahwah, NJ: Lawrence Erlbaum.
- Lynch, M. (2000). Against reflexivity as an academic virtue and source of privileged knowledge. *Theory, Culture and Society*, 17(3), 26–54.
- Lyotard, J.-F. (1984). *The postmodern condition: A report on knowledge*. Minneapolis: University of Minnesota Press.
- Macaulay, A. C., Paradis, G., Potvin, L., Cross, E. J., Saad-Haddad, C., McComber, A., et al. (1997). The Kahnawake Schools Diabetes Prevention Project: Intervention, evaluation, and baseline results of a diabetes primary prevention program with a native community in Canada. *Preventive Medicine*, 26(6), 79–90.
- MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov models and other types of models for discrete-valued time series*. London: Chapman & Hall.
- MacDonald, R. R. (2002). The incompleteness of probability models and the resultant implications for theories of statistical inference. *Understanding Statistics*, 1(3), 167–189.
- MacDougall, D. (2001). Renewing ethnographic film: Is digital video changing the genre? *Anthropology Today*, 17(3), 15–21.
- Machlin, S. R., & Taylor, A. K. (2000). *Design, methods, and field results of the 1996 Medical Expenditure Panel Survey Medical Provider Component* (MEPS Methodology Report No. 9, AHRQ Pub. No. 00–0028). Rockville, MD: Agency for Healthcare Research and Quality.

- Mackay, A. L. (1977). *Scientific quotations*. New York: Crane, Russak.
- Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford, UK: Oxford University Press.
- Mackintosh, N. J. (1995). *Cyril Burt: Fraud or framed?* Oxford, UK: Oxford University Press.
- Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, 28, 143–166.
- Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.
- Maddala, G. S. (1992). *Introduction to econometrics*. Englewood Cliffs, NJ: Prentice Hall.
- Madden, D. (2000). Towards a broader explanation of male-female wage differences. *Applied Economics Letters*, 7, 765–770.
- Madjar, I. (2001). The lived experience of pain in the context of clinical practice. In S. K. Toombs (Ed.), *Handbook of phenomenology and medicine* (pp. 263–277). Boston: Kluwer Academic.
- Magidson, J. (1993). The CHAID approach to segmentation modeling: CHi-squared Automatic Interaction Detection. In R. Bagozzi (Ed.), *Handbook of marketing research* (pp. 118–159). London: Blackwell.
- Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology*, 31, 223–264.
- Magnusson, D. (1967). *Test theory*. Reading, MA: Addison-Wesley.
- Mahoney, J. (2000). Path dependence in historical sociology. *Theory and Society*, 29, 507–548.
- Maines, D. R. (1992). Theorizing movement in an urban transportation system by use of the constant comparative method in field research. *Social Science Journal*, 29, 283–292.
- Maines, D. R. (2001). *The faultline of consciousness: A view of interactionism in sociology*. New York: Aldine De Gruyter.
- Makridakis, S., & Hibbon, M. (2000). The M3-competition: Results, conclusions, and implications. *International Journal of Forecasting*, 16, 451–476.
- Malinowski, B. (1989). *A diary in the strict sense of the term*. Stanford, CA: Stanford University Press.
- Malinowski, B. (1922). *Argonauts of the Western Pacific*. London: Routledge & Kegan Paul.
- Manski, C. F. (1995). *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.
- Manton, K. G., & XiLiang, G. (2003). *Variation in disability decline and Medicare expenditures*. Working Monograph, Center for Demographic Studies.
- Manton, K. G., & Stallard, E. (1988). *Chronic disease modelling: Vol. 2. Mathematics in medicine*. New York: Oxford University Press.
- Manton, K. G., Singer, B., & Woodbury, M. A. (1992). Some issues in the quantitative characterization of heterogeneous populations. In J. Trussell, R. Hankinson, & J. Tilton (Eds.), *Demographic application of event history analysis* (pp. 9–37). Oxford, UK: Clarendon.
- Manton, K. G., Woodbury, M. A., & Tolley, H. D. (1994). *Statistical applications using fuzzy sets*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley.
- Marcus, G., & Fischer, M. (1986). *Anthropology as cultural critique*. Chicago: University of Chicago Press.
- Marcuse, H. (1964). *One dimensional man*. Boston: Beacon.
- Marcuse, H. (1968). *One dimensional man: The ideology of industrial society*. London: Sphere.
- Mare, R. D. (1994). Discrete-time bivariate hazards with unobserved heterogeneity: A partially observed contingency table approach. In P. V. Marsden (Ed.), *Sociological methodology 1984* (pp. 341–383). Cambridge, MA: Blackwell.
- Markham, A. N. (1998). *Life online: Researching real experience in virtual space*. Walnut Creek, CA: AltaMira.
- Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12, 591–612.
- Marquis, K. (1984). Record checks for sample surveys. In T. Jabine, E. Loftus, M. Straf et al. (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines*. Washington, DC: National Academy Press.
- Marsh, C. (1982). *The survey method: The contribution of surveys to sociological explanation*. London: Allen & Unwin.
- Marsh, C. (1988). *Exploring data: An introduction to data analysis for social scientists*. Cambridge, UK: Polity.
- Marsh, L. C., & Cormier, D. R. (2001). *Spline regression models*. Thousand Oaks, CA: Sage.
- Marshall, P. A. (1992). Research ethics in applied anthropology. *IRB: A Review of Human Subjects Research*, 14(6), 1–5.
- Martens, H., & Naes, T. (1989). *Multivariate calibration*. London: Wiley.
- Martin, J. A., Hamilton, B. E., Ventura, S. J., Menacker, F., & Park, M. M. (2002, February 12). *Births: Final data for 2000* (National Vital Statistics Report, Vol. 50, No. 5). Hyattsville, MD: National Center for Health Statistics.
- Martin, K. (1998). *When a baby dies of AIDS: The parents' grief and searching for a reason*. Edmonton, Canada: Qual Institute Press.
- Mason, J. (1996). *Qualitative researching*. London: Sage.
- Mason, J. (2002). *Qualitative researching* (2nd ed.). London: Sage.
- Mason, K. O., Mason, W. M., Winsborough, H. H., & Poole, W. K. (1973). Some methodological issues in the cohort analysis of archival data. *American Sociological Review*, 38, 242–258.
- Mathiowetz, N., & McGonagle, K. (2000). An assessment of the current state of dependent interviewing in household surveys. *Journal of Official Statistics*, 16(4), 401–418.
- Matza, D. (1969). *Becoming deviant*. Englewood Cliffs, NJ: Prentice-Hall.
- Maxim, P. S. (1999). *Quantitative research methods in the social sciences*. New York: Oxford University Press.

- Maxwell, J. A. (1996). *Qualitative research design*. Thousand Oaks, CA: Sage.
- Maxwell, S. E., & Delaney, H. D. (1990). *Design experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.
- May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, 26, 459–467.
- Maynard, D. W. (1984). *Inside plea bargaining: The language of negotiation*. New York: Plenum.
- McAdams, D. P., & Constantian, C. A. (1983). Intimacy and affiliation motives in daily living: An experience sampling analysis. *Journal of Personality and Social Psychology*, 45, 851–861.
- McCleary, R., & Hay, R. A., Jr. (1980). *Applied time series analysis for the social sciences*. London: Sage.
- McCloskey, D. N. (1986). *The rhetoric of economics*. Madison: University of Wisconsin Press.
- McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42, 109–142.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.
- McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear and mixed models*. New York: John Wiley.
- McCullough, B. D., & Vinod, H. D. (1997). The numerical reliability of econometric software. *Journal of Economic Literature*, 37(2), 633–665.
- McDonald, J. F., & Moffitt, R. A. (1980). The uses of Tobit analysis. *Review of Economics and Statistics*, 62, 318–321.
- McDowall, D., McCleary, R., Meindinger, E. E., & Hay, R. A., Jr. (1980). *Interrupted time series analysis*. Beverly Hills, CA: Sage.
- McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of econometrics* (pp. 105–142). New York: Academic Press.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46.
- McGrew, W. (1972). *An ethological study of children's behavior*. New York: Academic Press.
- McGuigan, J. (1992). *Cultural populism*. London: Routledge.
- McGuinness, M., & Sapsford, R. (2002). *Middlesbrough Citizens' Advice Bureau: Report for 2001*. Middlesbrough, UK: University of Teesside, School of Social Sciences.
- McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. (1996). Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage*, 3, 143–157.
- McIver, J. P., & Carmines, E. G. (1980). *Unidimensional scaling*. Beverly Hills, CA: Sage.
- McKean, K. (1987, January). The orderly pursuit of pure disorder. *Discover*, pp. 72–81.
- McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, 103–120.
- McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and application to clustering*. New York: Marcel Dekker.
- McSweeney, A. J. (1978). Effects of response cost on the behavior of a million persons: Charging for directory assistance in Cincinnati. *Journal of Applied Behavior Analysis*, 11, 47–51.
- Meacham, S. J. (1998). Threads of a new language: A response to Eisenhart's "On the subject of interpretive reviews." *Review of Educational Research*, 68(4), 401–407.
- Mead, M. (1928). *Coming of age in Samoa*. New York: William Morrow.
- Meadows, L. M., & Dodendorf, D. M. (1999). Data management and interpretation: Using computers to assist. In B. F. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (pp. 195–218). Thousand Oaks, CA: Sage.
- Melton, G. B., Levine, R. J., Koocher, G. P., Rosenthal, R., & Thompson, W. C. (1988). Community consultation in socially sensitive research: Lessons from clinical trials on treatments for AIDS. *American Psychologist*, 43, 573–581.
- Menard, S. (1995). *Applied logistic regression analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–106). Thousand Oaks, CA: Sage.
- Menard, S. (2002). *Applied logistic regression analysis* (2nd ed., Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–106). Thousand Oaks, CA: Sage.
- Menard, S. (2002). *Longitudinal research* (2nd ed.). Thousand Oaks, CA: Sage.
- Menard, S. (in press). *Logistic regression*. Thousand Oaks, CA: Sage.
- Menard, S., & Elliott, D. S. (1990). Longitudinal and cross-sectional data collection and analysis in the study of crime and delinquency. *Justice Quarterly*, 7, 11–55.
- Menchú, R. (1984). *I, Rigoberta Menchú: An Indian woman in Guatemala*. (Ed. and intro. Elisabeth Burgos-Debray, Trans. Ann Wright.) London: Verso.
- Mendel, G. (1866). Versuche über Pflanzen-Hybriden [Experiments in plant hybridization]. In *Verhandlungen des naturforschenden Vereins* [Proceedings of the Natural History Society]. Available in both the original German and the English translation at www.mendelweb.org
- Merkle, D. M., & Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 243–258). New York: Wiley.
- Merleau-Ponty, M. (1962). *Phenomenology of perception*. London: Routledge Kegan Paul.
- Merton, R. K. (1949). *Social theory and social structure*. New York: Free Press.
- Merton, R. K. (1968). *Social theory and social structure* (3rd ed.). New York: Free Press.
- Merton, R. K. (1987). The focused interview and focus groups: Continuities and discontinuities. *Public Opinion Quarterly*, 51, 550–566.

- Merton, R. K., & Kendall, P. L. (1946). The focused interview. *American Journal of Sociology*, 51(6), 514–557.
- Merton, R. K., Fiske, M., & Kendall, P. L. (1990). *The focused interview: A manual of problems and procedures* (2nd ed.). New York: Free Press.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education.
- Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (6th ed., p. 1487–1495). New York: Macmillan.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1092.
- Meulman, J. J. (1998). Book review of W. J. Krzanowski, & F. H. C. Marriott, *Multivariate Analysis. Part I. Distributions, Ordinations, and Inference*, London: Edward Arnold, 1994. *Journal of Classification*, 15, 287–293.
- Meulman, J. J., & Heiser, W. J. (2000). *Categories*. Chicago: SPSS.
- Michel, Y., & Haight, B. K. (1996). Using the Solomon four design. *Nursing Research*, 45(6), 367–369.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.
- Milgram, S. (1974). *Obedience to authority*. London: Tavistock.
- Milgram, S. (1974). *Obedience to authority: An experimental view*. New York: Harper & Row.
- Mill, J. S. (1947). *A system of logic*. London: Longman, Green. (Originally published in 1879)
- Mill, J. S. (1956). *A system of logic, ratiocinative and inductive*. London: Longmans, Green & Company. (Original work published 1843)
- Mill, J. S. (1973). *A system of logic*. Toronto: University of Toronto Press. (Original work published 1865)
- Mill, J. S. (1974). *Collected works of John Stuart Mill: Vol. 8. A system of logic* (Books 4–6). Toronto: University of Toronto Press.
- Millar, A., Simeone, R. S., & Carnevale, J. T. (2001). Logic models: A system tool for performance management. *Evaluation and Program Planning*, 24, 73–81.
- Miller, A. G. (1986). *The obedience experiments: A case study of controversy in social science*. New York: Praeger.
- Miller, A. G. (1995). Constructions of the obedience experiments: A focus upon domains of relevance. *Journal of Social Issues*, 51, 33–53.
- Miller, A. J. (1990). *Subset selection in regression*. New York: Chapman Hall.
- Miller, D. (1998). Writing and retelling multiple ethnographic tales of a soup kitchen for the homeless. *Qualitative Inquiry*, 4, 469–492.
- Miller, L., Rustin, M., Rustin, M., & Shuttleworth, J. (1989). *Closely observed infants*. London: Duckworth.
- Miller, R. G. (1981). *Simultaneous statistical inference* (2nd ed.). New York: Springer-Verlag.
- Miller, R. W. (1987). *Fact and method: Explanation, confirmation and reality in the natural and the social sciences*. Princeton, NJ: Princeton University Press.
- Miller, W. L., & Crabtree, B. F. (1999). Clinical research: A multimethod typology and qualitative roadmap. In B. F. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (2nd ed., pp. 3–32). Thousand Oaks, CA: Sage.
- Millett, K. (1970). *Sexual politics*. Garden City, NY: Doubleday.
- Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms. *Psychometrika*, 44(3), 343–346.
- Milliken, G. A., & Johnson, D. A. (1984). *Analysis of messy data: Vol. 1. Designed experiments*. New York: Van Nostrand Reinhold.
- Milliken, G. A., & Johnson, D. E. (1992). *Analysis of messy data: Vol. 1. Designed experiments*. London: Chapman & Hall.
- Mills, C. W. (1959). *The sociological imagination*. New York: Oxford University Press.
- Mills, T. C. (1990). *Time series techniques for economists*. Cambridge, UK: Cambridge University Press.
- Millward, L. J. (2000). Focus groups. In G. M. Breakwell, C. Fife-Schaw, & S. Hammond (Eds.), *Research methods in psychology* (2nd ed., pp. 303–324). London: Sage.
- Milner, A. (1996). *Literature, culture and society*. London: UCL Press.
- Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy*, 66, 281–302.
- Mincer, J. (1974). *Schooling, experience and earnings*. New York: Columbia University Press.
- Mishler, E. G. (1986). *Research interviewing: Context and narrative*. Cambridge, MA: Harvard University Press.
- Mishler, E. G. (1991). Representing discourse: The rhetoric of transcription. *Journal of Narrative and Life History*, 1(4), 255–280.
- Mishler, E. G. (1995). Models of narrative analysis: A typology. *Journal of Narrative and Life History*, 5(2), 87–123.
- Mitchell, J. (1983). Case and situation analysis. *Sociological Review*, 31(2), 186–211.
- Mitchell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, 100, 398–407.
- Mittelhammer, R. C., Judge, G. G., & Miller, D. J. (2000). *Econometric foundations*. New York: Cambridge University Press.
- Mohr, D. C., Likosky, W., Bertagnolli, A., Goodkin, D. E., Van der Wende, J., Dwyer, P., et al. (2000). Telephone

- administered cognitive-behavioral therapy for the treatment of depressive symptoms in multiple sclerosis. *Journal of Consulting and Clinical Psychology*, 68(2), 356–361.
- Mohr, L. B. (1990). *Understanding significance testing* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–073). Newbury Park, CA: Sage.
- Monge, P., & Contractor, N. (2003). *Theories of communication networks*. New York: Oxford University Press.
- Monk-Turner, E., & Turner, C. G. (2001). Sex differentials in the South Korean labour market. *Feminist Economics*, 7(1), 63–78.
- Mood, A. M., & Graybill, F. A. (1963). *Introduction to the theory of statistics* (2nd ed.). New York: McGraw-Hill.
- Mooney, C. Z. (1997). *Monte Carlo simulation*. Thousand Oaks, CA: Sage.
- Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Thousand Oaks, CA: Sage.
- Moore, D. S. (1997). *Statistics: Concepts and controversies* (4th ed.). New York: W. H. Freeman.
- Moore, D. S., & McCabe, G. P. (1998). *Introduction to the practice of statistics*. New York: W. H. Freeman.
- Moore, D. W. (1999, June/July). Daily tracking polls: Too much “noise” or revealed insights? *Public Perspective*, pp. 27–31.
- Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology*, 65, 131–149.
- Moreno, L. (1994). Frailty selection in bivariate survival models: A cautionary note. *Mathematical Population Studies*, 4, 225–233.
- Morgan, D. L. (1988). *Focus groups as qualitative research*. Newbury Park, CA: Sage.
- Morgan, D. L. (1993). *Successful focus groups: Advancing the state of the art*. London: Sage.
- Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: Applications for health research. *Qualitative Health Research*, 8, 362–376.
- Morgan, D. L., & Krueger, R. A. (Eds.). (1998). *Focus group kit*. Thousand Oaks, CA: Sage.
- Morley, D. (1980). *The “Nationwide” audience*. London: British Film Institute.
- Morley, D. (1986). *Family television—Cultural power and domestic leisure*. London: British Film Institute.
- Morris, M., & Western, B. (1999). Inequality in earnings at the close of the 20th century. *Annual Review of Sociology*, 25, 623–657.
- Morrison, D. E., & Henkel, R. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.
- Morrow, R. A. (with Brown, D. D.). (1994). *Critical theory and methodology*. London: Sage.
- Morse, J. M. (1989). Strategies for sampling. In J. Morse (Ed.), *Qualitative nursing research: A contemporary dialogue* (pp. 117–131). Rockville, MD: Aspen.
- Morse, J. M. (1997). Considering theory derived from qualitative research. In J. Morse (Ed.), *Completing a qualitative project: Details and dialogue* (pp. 2163–2188). Thousand Oaks, CA: Sage.
- Morse, J. M., & Richards, L. (2002). *Read me first for a user’s guide to qualitative methods*. Thousand Oaks, CA: Sage.
- Morse, J. M., Swanson, J. M., & Kuzel, A. J. (Eds.). (2001). *The nature of qualitative evidence*. Thousand Oaks, CA: Sage.
- Morton, Rebecca B. (2000). *Methods and models: A guide to the empirical analysis of formal models in political science*. New York: Cambridge University Press.
- Moser, C. A., & Kalton, G. (1971). *Survey methods in social investigation* (2nd ed.). Aldershot, UK: Gower.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.
- Mounce, H. O. (1997). *The two pragmatisms: From Peirce to Rorty*. London: Routledge and Kegan Paul.
- Moustakas, C. (1990). *Heuristic research: Design, methodology, and applications*. Newbury Park, CA: Sage.
- Moustakas, C. (1994). *Phenomenological research methods*. Thousand Oaks, CA: Sage.
- Moyé, L. A. (2000). *Statistical reasoning in medicine*. New York: Springer-Verlag.
- Mruck, K., Corti, L., Kluge, S. & Opitz, D. (Eds.). (2000, December). Text archive: Re-analysis. *Forum Qualitative Sozialforschung* [Forum: Qualitative Social Research] [Online Journal], 1(3). Available: <http://qualitative-research.net/fqs/fqs-eng.htm>
- Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, 56, 63–75.
- Mueller, D. J. (1986). *Measuring social attitudes: A handbook for researchers and practitioners*. New York: Teachers College, Columbia University.
- Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS*. New York: Springer.
- Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.
- Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research*, 22, 267–305.
- Murphy, M. (1990). Minimising attrition in longitudinal studies: Means or end? In D. Magnusson & L. R. Bergman (Eds.), *Data quality in longitudinal research* (pp. 148–156). Cambridge, UK: Cambridge University Press.
- Murphy, M. K., Black, N. A., Lamping, D. L., McKee, C. M., Sanderson, C. F. B., Ashkam, J., & Marteau, T. (1998). Consensus development methods, and their use in clinical guideline development [Entire issue]. *Health Technology Assessment*, 2(3).
- Musch, J., & Reips, U. (2000). A brief history of Web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 61–88). San Diego, CA: Academic Press.

- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115–132.
- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29(1), 81–117.
- Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 51, 431–462.
- Myers, J. L., & Well, A. D. (1995). *Research design and statistical analysis*. Hillsdale, NJ: Lawrence Erlbaum.
- Naes, T., & Risvik, E. (Eds.). (1996). *Multivariate analysis of data in sensory science*. New York: Elsevier.
- Nagel, E. (1979). *The structure of science: Problems in the logic of scientific explanation*. Indianapolis, IN: Hackett.
- Nagler, J. (1995). Coding style and good computing practices. *Political Methodologist*, 6, 2–8.
- Namboodiri, K. (1984). *Matrix algebra: An introduction*. Beverly Hills, CA: Sage.
- Narayan, K. (1993). How native is the “native” anthropologist? *American Anthropologist*, 95, 671–686.
- National Center for Health Statistics. (2003). *United States life tables, 2000*. Retrieved from www.cdc.gov/nchs/data/lt2000.pdf.
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research* (DHEW Publication No. (OS) 78–0012). Washington, DC: Government Printing Office.
- National Institutes of Health. (2003, July 22). *Certificates of confidentiality*. Retrieved from <http://grants1.nih.gov/grants/policy/coc/index.htm>
- National Research Council. (2001). *The 2000 census: Interim assessment*. Washington, DC: National Academy Press.
- Naughton, J. (1984). *Soft systems analysis: An introductory guide*. Milton Keynes, UK: Open University Press.
- Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. San Francisco: W. H. Freeman.
- Nelder, J. A., & Lane, P. W. (1995). The computer analysis of factorial experiments: In memoriam—Frank Yates. *The American Statistician*, 49, 382–385.
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135, 370–384.
- Nelson, J. S., Megill, A., & McCloskey, D. N. (Eds.). (1987). *The rhetoric of the human sciences*. Madison: University of Wisconsin Press.
- Nesbary, D. K. (2000). *Survey research and the World Wide Web*. Boston, MA: Allyn and Bacon.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear regression models* (3rd ed.). Chicago: Irwin.
- Neter, J., Wasserman, W., & Kutner, M. H. (1989). *Applied linear regression models* (2nd ed.). Boston: Irwin.
- Neter, J., Wasserman, W., & Kutner, M. (1990). *Applied linear statistical models* (3rd ed.). Homewood, IL: Irwin.
- Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, 22, 287–293.
- New, C. (1996). *Agency, health and social survival*. London: Taylor & Francis.
- Newbold, P., & Bos, T. (1985). *Stochastic parameter regression models* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–051). Beverly Hills, CA: Sage.
- Newbold, P., & Bos, T. (1994). *Introductory business forecasting* (2nd ed.). Cincinnati, OH: South-Western.
- Newby, H. (1977). In the field: Reflections on a study of Suffolk farm workers. In C. Bell & H. Newby (Eds.), *Doing sociological research* (pp. 108–129). London: Allen and Unwin.
- Newell, A., & Simon, H. A. (1961). Computer simulation of human thinking. *Science*, 134, 2011–2017.
- Newell, A., & Simon, H. A. (1971). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.
- Newell, R. W. (1986). *Objectivity, empiricism and truth*. London: Routledge and Kegan Paul.
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55, 703–708.
- Newton, I. (1952). *Mathematical principles of natural philosophy*. Chicago: Britannica. (Originally published in 1686)
- Newton, R. R., & Rudestam, K. E. (1999). *Your statistical consultant: Answers to your data analysis questions*. Thousand Oaks, CA: Sage.
- NHS Centre for Reviews and Dissemination. (2001). *Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews* (Report number 4, 2nd ed.). York, UK: CRD.
- Nicholson, L. (Ed.). (1990). *Feminism/postmodernism*. New York: Routledge Kegan Paul.
- Nida, E. A. (1975). *Componential analysis of meaning: An introduction to semantic structures*. The Hague, The Netherlands: Mouton.
- Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.
- Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, N.J.: Lawrence Erlbaum.
- Nishisato, S. (1996). Gleaning in the field of dual scaling. *Psychometrika*, 61, 559–599.
- Nishisato, S., & Clavel, J. G. (2003). A note on between-set distances in dual scaling and correspondence analysis. *Behaviormetrika*, 30, 87–98.
- NIST/SEMATECH. (2003). *E-handbook of statistical methods*. Retrieved from <http://www.itl.nist.gov/div898/handbook/eda>
- Noblit, G. W., & Hare, R. D. (1988). *Meta-ethnography: Synthesizing qualitative studies*. Newbury Park, CA: Sage.
- Nocedal, J., & Wright, S. (1999). *Numerical optimization*. New York: Springer-Verlag.
- Noldus, L. P. J. J., Trienes, R. J. H., Hendriksen, A. H. M., Jansen, H., & Jansen, R. G. (2000). The Observer

- Video-Pro: New software for the collection, management, and presentation of time-structured data from videotapes and digital media files. *Behavior Research Methods, Instruments & Computers*, 32, 197–206.
- Norpoth, H. (1995). Is Clinton doomed? An early forecast for 1996. *Political Science & Politics*, 27, 201–206.
- Norton, H. W. (1939). The 7 × 7 squares. *Annals of Eugenics*, 9, 269–307.
- Norusis, M. (1990). *SPSS/PC+ Statistics 4.0*. Chicago: SPSS.
- Norusis, M. J. (1990). *SPSS base system user's guide*. Chicago: SPSS.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-research: Ethics, security, design, and control in psychological research on the Internet. *Journal of Social Issues*, 58, 161–176.
- Nowak, A., & Latané, B. (1994). Simulating the emergence of social order from individual behaviour. In N. Gilbert & J. Doran (Eds.), *Simulating societies: The computer simulation of social phenomena* (pp. 63–84). London: UCL.
- Noymer, A. (2001). The transmission and persistence of “urban legends”: Sociological application of age-structured epidemic models. *Journal of Mathematical Sociology*, 25, 299–323.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Nutley, S., & Webb, J. (2000). Evidence and the policy process. In H. Davies, S. Nutley, & P. Smith (Eds.), *What works? Evidence-based policy and practice in public services* (pp. 13–41). Bristol, UK: Policy Press.
- O'Brien, R. M. (2000). Age-period-cohort-characteristic models. *Social Science Research*, 29, 123–139.
- O'Hear, A. (1989). *An introduction to the philosophy of science*. Oxford, UK: Clarendon.
- O'Hear, A. (Ed.). (1996). *Verstehen and humane understanding*. Supplement to *Philosophy*, Royal Institute of Philosophy Supplement 41. Cambridge, UK: Cambridge University Press.
- O'Rourke, D., & Blair, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research*, 20, 428–432.
- Oakley, A. (2000). *Experiments in knowing: Gender and method in the social sciences*. Cambridge, UK: Polity.
- Ohnuki-Tierney, E. (1984). “Native” anthropologists. *American Ethnologist*, 11, 584–586.
- Oksenberg, L., Cannell, C., & Kalton, G. (1991). New strategies for pretesting survey questions. *Journal of Official Statistics*, 7(3), 349–365.
- Olmsted, P., & Weikart, D. P. (Eds.). (1994). *Families speak: Early childhood care and education in 11 countries*. Ypsilanti, MI: High/Scope Press.
- Olsen W. K. (1993) Random samples and repeat surveys in South India. In S. Devereux & J. Hoddinott (Eds.) *Fieldwork in Developing Countries*, (pp. 57–72). Boulder, CO: Lynne Rienner.
- Olsen, W. K. (1999). *Path analysis for the study of farming and micro-enterprise* (Bradford Development Paper No. 3). Bradford, UK: University of Bradford, Development and Project Planning Centre.
- Olsen, W. K., Warde, A., & Martens, L. (2000). Social differentiation and the market for eating out in the UK. *International Journal of Hospitality Management*, 19(2), 173–190.
- Oneal, J. R., & Russett, B. (1997). The classical liberals were right: Democracy, interdependence, and conflict, 1950–1985. *International Studies Quarterly*, 41, 267–294.
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776–783.
- Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 143–179). New York: Academic Press.
- Orr, L. L. (1999). *Social experiments: Evaluating public programs with experimental methods*. Thousand Oaks, CA: Sage.
- Orwin, R. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8, 157–159.
- Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, 5, 343–355.
- Osgood, C. E. (1952). The nature of and measurement of meaning. *Psychological Bulletin*, 49, 197–237.
- Osgood, C. E., & Tzeng, O. C. S. (Eds.). (1990). *Language, meaning, and culture: The selected papers of C. E. Osgood*. New York: Praeger.
- Osgood, C. E., Suci, C. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.
- Ostrom, C. W., Jr. (1990). *Time series analysis*. Thousand Oaks, CA: Sage.
- Otis, D. L., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, 62, 1–135.
- Outhwaite, W. (1975). *Understanding social life: The method called verstehen*. London: Allen & Unwin.
- Outhwaite, W. (1987). *New philosophies of social science: Realism, hermeneutics and critical theory*. London: Macmillan.
- Paccagnella, L. (1997). Getting the seats of your pants dirty: Strategies for ethnographic research on virtual communities. *Journal of Computer Mediated Communication*, 3(1). Retrieved from <http://www.ascusc.org/jcmc/vol3/issue1/paccagnella.html>
- Paechter, C. (1996). Power, knowledge and the confessional in qualitative research. *Discourse: Studies in the Politics of Education*, 17(1), 75–84.
- Page, B. I., & Shapiro, R. Y. (1992). *The rational public*. Chicago: University of Chicago Press.
- Page, S. (2000). The lost art of unobtrusive measures. *Journal of Applied Social Psychology*, 30(10), 2126–2128.

- Pagès, J., & Tenenhaus, M. (2001). Multiple factor analysis combined with PLS path modeling: Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgments. *Chemometrics and Intelligent Laboratory Systems*, 58, 261–273.
- Palmer, R. E. (1969). *Hermeneutics: Interpretation theory of Schleiermacher, Dilthey, Heidegger, and Gadamer*. Evanston, IL: Northwestern University Press.
- Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.
- Pankratz, L. (1979). Symptom validity testing and symptom retraining: Procedures for the assessment and treatment of functional sensory deficits. *Journal of Consulting and Clinical Psychology*, 47, 409–410.
- Papineau, D. (1993). *Philosophical naturalism*. Oxford, UK: Basil Blackwell.
- Park, A., & Jowell, R. (1997). *Consistencies and differences in a cross-national survey*. London: SCPR.
- Parker, I. (1988). Deconstructing accounts. In C. Antaki (Ed.), *Analysing everyday explanation: A casebook of methods* (pp. 184–198). London: Sage.
- Parkin, F. (1973). *Class inequality and political order*. London: Paladin.
- Parkinson, B., & Manstead, A. S. R. (1993). Making sense of emotion in stories and social life. *Cognition and Emotion*, 7, 295–323.
- Parsons, T. (1968). Émile Durkheim. In D. L. Sills (Ed.), *International encyclopedia of the social sciences*, Vol. 4. New York: Macmillan.
- Parsons, T. (1968). *The structure of social action* (3rd ed.). New York: Free Press.
- Passmore, J. (1967). Logical positivism. In P. Edwards (Ed.), *The encyclopedia of philosophy* (Vol. 5, pp. 52–57). New York: Macmillan.
- Paterson, B. L., Thorne, S. E., Canam, C., & Jillings, C. (2001). *Meta-study of qualitative health research: A practical guide to meta-analysis and meta-synthesis*. Thousand Oaks, CA: Sage.
- Patrick, C. J. (1994). Emotion and psychopathy: Startling new insights. *Psychophysiology*, 31, 415–428.
- Pattison, P. E. (1993). *Algebraic models for social networks*. New York: Cambridge University Press.
- Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.
- Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.
- Patton, M. Q. (1999). *Grand Canyon celebration: A father-son journey of discovery*. Amherst, NY: Prometheus.
- Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598–609.
- Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 6(2), 559–572.
- Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, 3, 1243–1246.
- Peat, J., Mellis, C., Williams, K., & Xuan W. (2002). *Health science research: A handbook of quantitative methods*. London: Sage.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: Holt, Rinehart and Winston.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). New York: Wadsworth.
- Peirce, C. S. (1931–1958). *Collected papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press.
- Pelusi, J. (1997). The lived experience of surviving breast cancer. *Oncology Nursing Forum*, 24(8), 1343–1353.
- Perks, R., & Thomson, A. (Eds.). (1998). *The oral history reader*. London: Routledge.
- Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, 21, 381–391.
- Pettoufrezzo, A. J. (1978). *Matrices and transformations*. New York: Dover.
- Pfungst, O. (1965). *Clever Hans: The horse of Mr. von Osten*. New York: Holt, Rinehart, & Winston. (Original work published 1911)
- Phillips, D. (1976). *Holistic thought in social science*. Stanford, CA: Stanford University Press.
- Phillips, N., & Brown, J. (1993). Analyzing communication in and around organizations: A critical hermeneutic approach. *Academy of Management Journal*, 36(6), 1547–1576.
- Phillips, P. C. B. (1991). Optimal inference in cointegrated systems. *Econometrica*, 59, 283–306.
- Piaget, J. (1929). *The child's conception of the world*. New York: Harcourt, Brace Jovanovich.
- Piaget, J. (1955). *The construction of reality in the child*. New York: Routledge Kegan Paul. (Original work published 1937)
- Pickering, M. (2001). *Stereotyping: The politics of representation*. Basingstoke, UK: Palgrave.
- Pierce, J. L., Gardner, D. G., Cummings, L. L., & Dunham, R. B. (1989). Organization-based self-esteem: Construct definition, measurement, and validation. *Academy of Management Journal*, 32, 622–648.
- Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. *American Political Science Review*, 94, 251–267.
- Pike, K. L. (1967). *Language in relation to a unified theory of the structure of human behavior* (2nd ed.). The Hague: Mouton. (Original work published 1954)
- Pindyck, R. S., & Rubinfeld, D. L. (1991). *Econometric models and economic forecasts* (3rd ed.). New York: McGraw-Hill.

- Pindyck, R. S., & Rubinfeld, D. L. (1998). *Econometric models and economic forecasts* (4th ed.). New York: McGraw-Hill.
- Pink, S. (2001). *Doing visual ethnography: Images, media and representation in research*. London: Sage.
- Plackett, R. L. (1972). The discovery of the method of least squares. *Biometrika*, 59, 239–251.
- Plake, B. S., & Impara, J. C. (Eds.). (2001). *The fourteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
- Platt, J. (2002). The history of the interview. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research: Context and method* (pp. 33–54). Thousand Oaks, CA: Sage.
- Plessy v. Ferguson, 163 U.S. 537 (1896).
- Plewis, I. (1985). *Analysing change: Measurement and exploration using longitudinal data*. Chichester, UK: Wiley.
- Plummer, K. (2001). *Documents of life 2: An invitation to a critical humanism*. London: Sage.
- Poland, B. (2001). Transcription quality. In J. Gubrium & J. Holstein (Eds.), *Handbook of interview research*. Thousand Oaks, CA: Sage.
- Poland, B. D. (1995). Transcript quality as an aspect of rigor in qualitative research. *Qualitative Inquiry*, 1(3), 290–310.
- Pollock, K. H., Nichols, J. D., Brownie, C., & Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs*, 107, 1–97.
- Pope, C., & Mayes, N. (1995). Qualitative research: Reaching the parts other methods cannot reach: An introduction to qualitative methods in health and health services research. *British Medical Journal*, 311, 42–45.
- Popkin, S. L. (1991). *The reasoning voter*. Chicago: University of Chicago Press.
- Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.
- Popper, K. (1965). *Conjectures and refutations: The growth of scientific knowledge* (2nd ed.). New York: Basic Books.
- Popper, K. (1966). *The open society and its enemies* (5th ed., Vol. 1). London: Routledge. (Original work published 1945)
- Popper, K. (1972). *Objective knowledge: An evolutionary approach*. Oxford, UK: Clarendon.
- Popper, K. R. (1957). *The poverty of historicism*. Boston: Beacon.
- Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.
- Popper, K. R. (1961). *The poverty of historicism*. London: Routledge & Kegan Paul.
- Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. New York: Basic Books.
- Popper, K. R. (1965). *The logic of scientific discovery*. New York: Harper & Row.
- Popper, K. R. (1972). *Conjectures and refutation*. London: Routledge & Kegan Paul.
- Popper, K. E. (1986). *The poverty of historicism*. London: ARK.
- Porter, S. (1993). Critical realist ethnography: The case of racism and professionalism in a medical setting. *Sociology*, 27, 591–609.
- Potter, J. (1996). *Representing reality: Discourse, rhetoric and social construction*. London: Sage.
- Potter, J., & Wetherell, M. (1987). *Discourse and social psychology: Beyond attitudes and behaviour*. London: Sage.
- Poundstone, W. (1988). *Labyrinths of reason: Paradoxes, puzzles and the frailty of knowledge*. New York: Anchor/Doubleday.
- Poundstone, W. (1992). *Prisoner's dilemma*. New York: Doubleday.
- Power, M. (1990). Modernism, postmodernism and organisation. In J. Hassard & D. Pym (Eds.), *The theory and philosophy of organisations*. London: Routledge Kegan Paul.
- Powers, D. A., & Xie, Y. (2000). *Statistical methods for categorical data analysis*. San Diego: Academic Press.
- Prais, S. J., & Winsten, C. B. (1954). *Trend estimators and serial correlation*. Chicago: Cowles Commission.
- Prendergast, C. (1995). *Cultural materialism: On Raymond Williams*. Minneapolis: University of Minnesota Press.
- Prentice, R. L., & Zhao, L. P. (1991). Estimating equations for parameters in mean and covariances of multivariate discrete and continuous responses. *Biometrics*, 47, 825–839.
- Preston, S. H., & Coale, A. J. (1982). Age structure, growth, attrition and accession: A new synthesis. *Population Index*, 48(2), 217–259.
- Preston, S. H., Heuveline, P., & Guillot, M. (2001). *Demography: Measuring and modeling population processes*. Oxford, UK: Basil Blackwell.
- Price, L. J. (1989). In the shadow of biomedicine: Self-medication in two Ecuadorian pharmacies. *Social Science and Medicine*, 28, 905–915.
- Price, L. J. (2002). Carrying out a structured observation. In M. V. Angrosino (Ed.), *Doing cultural anthropology* (pp. 107–114). Prospect Heights, IL: Waveland.
- Price, V. (1992). *Public opinion*. Newbury Park, CA: Sage.
- Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika*, 35, 73–78.
- Przeworski, A., & Teune, H. (1970). *The logic of comparative social inquiry*. New York: John Wiley.
- Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, 11, 18–84.
- Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2001). *GLLAMM manual technical report 2001/01*. London: King's College, University of London, Department of Biostatistics and Computing, Institute of Psychiatry. Retrieved from www.iop.kcl.ac.uk/IoP/Departments/BioComp/programs/manual.pdf
- Rachlin, H. (2000). *The science of self-control*. Cambridge, MA: Harvard University Press.
- Radest, H. B. (1996). *Humanism with a human face: Intimacy and the enlightenment*. London: Praeger.
- Radley, A., & Taylor, D. (2003). Remembering one's stay in hospital: A study in recovery, photography and forgetting.

- Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine*, 7(2), 129–159.
- Rae, D. W., & Yates, D. (1981). *Equalities*. Cambridge, MA: Harvard University Press.
- Raftery, A. E. (1996). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (Vol. 25, pp. 111–163). Oxford, UK: Basil Blackwell.
- Raftery, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 163–187). London: Chapman and Hall.
- Ragunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54–63.
- Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.
- Ragin, C. C. (1994). *Constructing social research*. Thousand Oaks, CA: Pine Forge Press.
- Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago: University of Chicago Press.
- Ragin, C. C., & Becker, H. S. (1992). *What is a case? Exploring the foundations of social inquiry*. New York: Cambridge University Press.
- Ramazanoglu, C., with Holland, J. (2002). *Feminist methodology*. London: Sage.
- Ramsay, J. O. (1977). Maximum likelihood estimation in MDS. *Psychometrika*, 42, 241–266.
- Ramsey, F. P. (1960). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays*. New York: Harcourt Brace.
- Random House. (1997). *Random House Webster's college dictionary*. New York: Author.
- Rao, P., & Griliches, Z. (1969). Small-sample properties of several two-stage regression methods in the context of auto-correlated errors. *Journal of the American Statistical Association*, 64, 253–272.
- Rapoport, A., & Chammah, A. M. (1965). *Prisoner's dilemma*. Ann Arbor: University of Michigan Press.
- Rässler, S. (2002). Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches. *Lecture Notes in Statistics*, 168. New York: Springer.
- Raudenbush, S. W., & Bryk, A. S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. *Review of Research in Education*, 15, 423–476.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1999) *Applied regression analysis: A research tool*. New York: Springer.
- Reason, R., & Brabury, H. (Eds.). (2001). *Handbook of action research: Participative inquiry and practice*. London: Sage.
- Redington, M., & Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in the Cognitive Sciences*, 1(7), 273–281.
- Reed-Danahay, D. (2001). Autobiography, intimacy and ethnography. In P. Atkinson, A. Coffey, S. Delamont, J. Lofland, & L. Lofland (Eds.), *Handbook of ethnography* (pp. 405–425). London: Sage.
- Reed-Danahay, D. E. (Ed.). (1997). *Auto/ethnography: Rewriting the self and the social*. Oxford, UK: Berg.
- Rees, P., Martin, D., & Williamson, P. (Eds.). (2002). *The census data system*. Chichester, UK: Wiley.
- Reichenbach, H. (1951). *The rise of scientific philosophy*. Berkeley: University of California Press.
- Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 190–222). New York: Cambridge University Press.
- Rengger, J. (1996). *Retreat from the modern: Humanism, post-modernism and the flight from modernist culture*. Exeter, UK: Bowerdean.
- Rescher, N. (1997). *Objectivity: The obligations of impersonal reason*. Notre Dame, IN: University of Notre Dame Press.
- Retherford, R. D., & Choe, M. K. (1993). *Statistical models for causal analysis*. New York: John Wiley.
- Rex, J. (1974). *Sociology and the demystification of the modern world*. London: Routledge & Kegan Paul.
- Reyment, R. A., & Jöreskog, K. G. (1993). *Applied factor analysis in the natural sciences*. Cambridge, UK: Cambridge University Press.
- Reynolds, H. T. (1984). *Analysis of nominal data* (2nd ed.). Beverly Hills, CA: Sage.
- Rhee, J. W. (1996). How polls drive campaign coverage. *Political Communication*, 13, 213–229.
- Rice, J. A. (1995). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury.
- Richards, T. J., & Richards, L. (1994). Using computers in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 445–462). Thousand Oaks, CA: Sage.
- Richardson, L. (1985). *The new other woman: Contemporary single women in affairs with married men*. London: Collier-Macmillan.
- Richardson, L. (1997). *Fields of play: Constructing an academic life*. New Brunswick, NJ: Rutgers University Press.
- Richardson, L. (2000). Writing: A method of inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 923–948). Thousand Oaks, CA: Sage.
- Richardson, L., & Lockridge, E. (in press). *Travels with Ernest: Crossing the Literary-Ethnographic Divide*. Walnut Creek, CA: AltaMira.
- Ricoeur, P. (1981). *Hermeneutics and the human sciences: Essays on language, action and interpretation*. New York: Cambridge University Press.
- Ricoeur, P. (1992). *Oneself as another*. Chicago: The University of Chicago Press.
- Ridgeway, C. L. (2001). Inequality, status, and the construction of status beliefs. In J. H. Turner (Ed.), *Handbook*

- of sociological theory (pp. 323–340). New York: Kluwer Academic/Plenum.
- Rieger, J. (1996). Photographing social change. *Visual Sociology*, 11(1), 5–49.
- Riessman, C. K. (2002). Doing justice: Positioning the interpreter in narrative work. In W. Patterson (Ed.), *Strategic narrative: New perspectives on the power of personal and cultural storytelling* (pp. 195–216). Lanham, MA: Lexington Books.
- Riessman, C. K. (2002). Positioning gender identity in narratives of infertility: South Indian women's lives in context. In M. C. Inhorn & F. van Balen (Eds.), *Infertility around the globe: New thinking on childlessness, gender, and reproductive technologies*. Berkeley: University of California Press.
- Riessman, C. K. (2003). Performing identities in illness narrative: Masculinity and multiple sclerosis. *Qualitative Research*, 3(1), 5–33.
- Rigdon, E. E. (1995). A necessary and sufficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research*, 30, 359–383.
- Riggs, P. J. (1992). *Whys and ways of science*. Melbourne, Australia: Melbourne University Press.
- Riker, W. H. (1980). Implications from the disequilibrium of majority rule for the study of institutions. *American Political Science Review*, 74, 432–446.
- Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge, UK: Cambridge University Press.
- Ritchie, J., & Lewis, J. (2003). *Qualitative research practice*. London: Sage.
- Ritzer, G. (1980). *Sociology: A multiple paradigm science*. Boston: Allyn & Bacon.
- Rivers, D., & Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39, 347–366.
- Robert, C. P. (2001). *The Bayesian choice: A decision theoretic motivation* (2nd ed.). New York: Springer-Verlag.
- Robert, C. P., & Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer-Verlag.
- Roberts, B. (2002). *Biographical research*. Buckingham, UK: Open University Press.
- Roberts, N., Andersen, D. F., Deal, R. M., Grant, M. S., & Shaffer, W. A. (1983). *Introduction to computer simulation: A system dynamics modeling approach*. Reading, MA: Addison-Wesley.
- Robey, B., Rutstein, S. O., & Morris, L. (1992). The reproductive revolution: New survey findings (Technical Report M-11). *Population Reports*. Baltimore, MD: Johns Hopkins University Center for Communication Programs.
- Robins, J. M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine*, 17, 269–302.
- Robinson, J. (1994). White women researching/representing “others”: From anti-apartheid to postcolonialism. In A. Blunt & G. Rose (Eds.), *Writing, women and space* (pp. 97–226). New York: Guilford.
- Robinson, J. P., & Godbey, G. (1997). *Time for life*. University Park: Pennsylvania State University Press.
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351–357.
- Robinson, W. S. (1951). The logical structure of analytic induction. *American Sociological Review*, 16, 812–818.
- Robson, C. (1993). *Real world research*. Oxford, UK: Blackwell.
- Rochford, E. B., Jr. (1992). On the politics of member validation: Taking findings back to Hare Krishna. In G. Miller & J. A. Holstein (Eds.), *Perspectives on social problems* (Vol. 3, pp. 99–116). Greenwich, CT: JAI.
- Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, 34, 441–456.
- Rodríguez, G. (1994). Statistical issues in the analysis of reproductive histories using hazard models. In K. L. Campbell & J. W. Wood (Eds.), *Human reproductive ecology: Interaction of environment, fertility and behavior* (pp. 266–279). New York: New York Academy of Sciences.
- Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the worker*. Cambridge, MA: Harvard University Press.
- Rogers, A. (1995). *Multiregional demography*. Chichester, UK: Wiley.
- Rohrer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, 63, 129–156.
- Romney, A. K., Weller, S. C., & Batchelder, W. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, 88, 313–338.
- Ronan, W. W., & Latham, G. P. (1974). The reliability and validity of the critical incident technique: A closer look. *Studies in Personnel Psychology*, 6(1), 33–64.
- Roncek, D. W. (1992). Learning more from Tobit coefficients. *American Sociological Review*, 57, 503–507.
- Rorty, R. (1991). *Objectivity, relativism, and truth: Philosophical papers volume I*. Cambridge, UK: Cambridge University Press.
- Rosaldo, R. (1989). *Culture & truth*. Boston: Beacon.
- Rose, G. (2001). *Visual methodologies*. London: Sage.
- Rosenau, P. M. (1992). *Post-modernism and the social sciences: Insights, inroads and intrusions*. Princeton, NJ: Princeton University Press.
- Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer-Verlag.
- Rosenbaum, P. R. (1998). Multivariate matching methods. *Encyclopedia of Statistical Sciences*, 2, 435–438.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rosenberg, A. (1988). *Philosophy of social science*. Oxford, UK: Clarendon.
- Rosenberg, M. (1968). *The logic of survey analysis*. New York: Basic Books.
- Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal &

- R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 279–349). New York: Academic Press.
- Rosenberg, S., & Kim, M. P. (1975). The method of sorting as a data-gathering method in multivariate research. *Multivariate Behavioral Research*, 10, 489–502.
- Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis*. New York: Wiley.
- Rosenhan, D. (1973). On being sane in insane places. *Science*, 179, 250–258.
- Rosenthal, G. (Ed.). (1998). *The Holocaust in three generations: Families of victims and persecutors of the Nazi regime*. London: Cassell.
- Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.
- Rosenthal, R. (1976). *Experimenter effects in behavioral research* (rev. ed.). New York: Irvington.
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86, 638–641.
- Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston.
- Rosenthal, R., & Jacobson, L. (1992). *Pygmalion in the classroom: Expanded edition*. New York: Irvington.
- Rosenthal, R., & Rosnow, R. L. (1975). *The volunteer subject*. New York: Wiley.
- Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.
- Rosenthal, R., & Rosnow, R. L. (Eds.). (1969). *Artifact in behavioral research*. New York: Academic Press.
- Rosenthal, R., & Rubin, D. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 3, 377–415.
- Rosenzweig, M. R., & Wolpin, K. I. (2000). Natural “natural experiments” in economics. *Journal of Economic Literature*, 38, 827–874.
- Rosenzweig, S. (1933). The experimental situation as a psychological problem. *Psychological Review*, 40, 337–354.
- Rosnow, R. L., & Rosenthal, R. (1970). Volunteer effects in behavioral research. In K. H. Craik, B. Kleinmuntz, R. L. Rosnow, R. Rosenthal, J. A. Cheyne, & R. H. Walters (Eds.), *New directions in psychology* (pp. 213–277). New York: Holt, Rinehart and Winston.
- Rosnow, R. L., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. New York: W. H. Freeman.
- Ross, S. (1999). *A first course in probability*. Upper Saddle River, NJ: Prentice Hall.
- Rossi, P. H. (1951). *The application of latent structure analysis to the study of social stratification*. Unpublished doctoral dissertation, Columbia University.
- Rossi, P. H. (1979). Vignette analysis: Uncovering the normative structure of complex judgments. In R. K. Merton, J. S. Coleman, & P. H. Rossi (Eds.), *Qualitative and quantitative social research: Papers in honor of Paul F. Lazarsfeld* (pp. 176–186). New York: Free Press.
- Rossi, P. H., & Anderson, A. B. (1982). The factorial survey approach: An introduction. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments: The factorial survey approach* (pp. 15–67). Beverly Hills, CA: Sage.
- Rossi, P. H., & Berk, R. A. (1985). Varieties of normative consensus. *American Sociological Review*, 50, 333–347.
- Rossi, P. H., Freeman, H. E., & Lipsey, M. (1998). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.
- Rossi, P. H., Sampson, W. A., Bose, C. E., Jasso, G., & Passel, J. (1974). Measuring household social standing. *Social Science Research*, 3, 169–190.
- Rothenberg, J., & Rothenberg, D. (Eds.). (1983). *Symposium of the whole: A range of discourse toward an ethnopoetics*. Berkeley: University of California Press.
- Rothman, K. J. (1977). Epidemiologic methods in clinical trials. *Cancer*, 39, 1771–1775.
- Rothman, K. J. (1986). *Modern epidemiology*. Boston: Little, Brown.
- Rothman, K. J., & Greenland, S. (1998). *Modern epidemiology* (2nd ed.). Philadelphia: Lippincott.
- Rousseau, J.-J. (1953) *The confessions*. Harmondsworth, UK: Penguin. (Original work published 1781)
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression & outlier detection*. New York: Wiley.
- Routledge, P. (1996). The third space as critical engagement. *Antipode*, 28(4), 399–419.
- Rowell, R. K. (1996). Partitioning predicted variance into constituent parts: How to conduct commonality analysis. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 4, pp. 33–44). Greenwich, CT: JAI.
- Royston, J. P. (1993). A toolkit for testing for non-normality in complete and censored samples. *The Statistician*, 42, 37–43.
- Rubenstein, A. (1998). *Modeling bounded rationality*. Cambridge: MIT Press.
- Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4, 87–95.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2, 169–188.
- Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366–374.
- Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo method*. New York: John Wiley.
- Rudas, T. (1997). *Odds ratios in the analysis of contingency tables*. Thousand Oaks, CA: Sage.
- Rummel, R. J. (1970). *Applied factor analysis*. Evanston, IL: Northwestern University Press.

- Ruskey, F. (2001). A survey of Venn diagrams. *The Electronic Journal of Combinatorics*, Dynamic Survey #5 [Online]. Retrieved from <http://www.combinatorics.org/Surveys/ds5/VennEJC.html>
- Russell, B. (1912). *Mysticism and logic*. London: Allen & Unwin.
- Russett, B. (1969). Inequality and instability: The relation of land tenure to politics. In D. Rowney & J. Graham (Eds.), *Quantitative history: Selected readings in the quantitative analysis of historical data* (pp. 356–367). Homewood, IL: Dorsey.
- Ryle, G. (1971). Thinking and reflecting. In G. Ryle, *Collected papers, Volume 2*. London: Hutchinson.
- Ryssevik, J., & Musgrave, S. (2001). The social science dream machine: Resource discovery, analysis, and delivery on the Web. *Social Science Computing Review*, 19(2), 163–174.
- Sackett, G. P. (1979). The lag sequential analysis of contingency and cyclicity in behavioral interaction research. In J. D. Osofsky (Ed.), *Handbook of infant development* (1st ed., pp. 623–649). New York: Wiley.
- Sacks, H. (1992). *Lectures on conversation* (2 vols.). Oxford, UK: Blackwell.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.
- Sade, A. (1951). An omission in Norton's list of 7×7 squares. *Annals of Mathematical Statistics*, 22, 306–307.
- Sale, J. E. M., Lohfeld, L. H., & Brazil, K. (2002). Revisiting the quantitative-qualitative debate: Implications for mixed-methods research. *Quality and Quantity*, 36, 43–53.
- Salem, A., & Mount, T. (1974). A convenient descriptive model of income distribution: The gamma density. *Econometrica*, 42, 1115–1127.
- Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.
- Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: Holt.
- Sandelowski, M., Docherty, S., & Emden, C. (1997). Qualitative metasynthesis: Issues and techniques. *Research in Nursing & Health*, 20, 365–371.
- Sankoff, D., & Kruskal, J. B. (1983). *Time warps, string edits, and macromolecules*. Reading, MA: Addison-Wesley.
- Sapir, E. (1951). Culture, genuine and spurious. In D. G. Mandelbaum (Ed.), *Selected writings of Edward Sapir in language, culture and personality* (pp. 308–331). Berkeley: University of California Press.
- Saris, W. E. (1991). *Computer-assisted interviewing*. Newbury Park, CA: Sage.
- Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. In C. C. Clogg (Ed.), *Sociological methodology 1987* (pp. 105–129). San Francisco: Jossey-Bass.
- Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.
- Saunders, M. N. K., & Cooper, S. A. (1993). *Understanding business statistics: An active learning approach*. London: DP Publications.
- Saunders, P. (1979). *Urban politics: A sociological interpretation*. London: Hutchinson.
- Saunders, P. T. (1980). *An introduction to catastrophe theory*. New York: Cambridge University Press.
- Saussure, F. (1966). *Course in general linguistics*. New York: McGraw-Hill. (Originally published in 1915)
- Saussure, F. de (1974). *Course in general linguistics*. London: Fontana/Collins.
- Sayer, A. (1992). *Method in social science: A realist approach* (2nd ed.). London: Routledge.
- Sayer, A. (1997). Essentialism, social constructionism and beyond. *Sociological Review*, 45, 453–487.
- Sayer, A. (2000). *Realism and social science*. London: Sage.
- Sayer, A. (2000). System, lifeworld and gender: Associational versus counterfactual thinking. *Sociology*, 34, 707–725.
- Sayer, A. (2001). Reply to Holmwood. *Sociology*, 35, 967–984.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Scheaffer, R. L., Mendenhall, W., & Ott, L. (1990) *Elementary survey sampling*. Boston: PWS-Kent.
- Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, 40, 87–104.
- Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.
- Schegloff, E. A. (1968). Sequencing in conversational openings. *American Anthropologist*, 70, 1075–1095.
- Schegloff, E. A. (1989). Harvey Sacks—lectures 1964–1965: An introduction/memoir. *Human Studies*, 12, 185–209.
- Schegloff, E. A. (1992). Repair after next turn: The last structurally provided defence of intersubjectivity in conversation. *American Journal of Sociology*, 97(5), 1295–1345.
- Schegloff, E.A. (1996). Confirming allusions: Toward an empirical account of action. *American Journal of Sociology*, 2, 161–216.
- Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1, 143–186.
- Schelling, T. C. (1978). *Micromotives and macrobehavior*. New York: Norton.
- Schensul, J. J., & LeCompte, M. D. (Eds.). (1999). *Ethnographer's toolkit*. Walnut Creek, CA: AltaMira.
- Schensul, S. L., Schensul, J. J., & LeCompte, M. D. (1999). *Essential ethnographic methods: Observations, interviews, and questionnaires*. Walnut Creek, CA: AltaMira.
- Schervish, M. J. (1996). *P-values: What they are and what are they not*. *American Statistician*, 50, 203–206.
- Schlesselman, J. J. (1982). *Case-control studies: Design, conduct, analysis*. New York: Oxford University Press.
- Schmidt, P. (1976). *Econometrics*. New York: Marcel Dekker.
- Schmidt, P., & Witte, A. D. (1988). *Predicting recidivism using survival models*. New York: Springer-Verlag.
- Schneider, H. (1986). *Truncated and censored samples from normal populations*. New York: Marcel Dekker.

- Schober, M., & Conrad, F. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576–602.
- Schoenberg, R. (1977). Dynamic models and cross-sectional data: The consequences of dynamic misspecification. *Social Science Research*, 6, 133–144.
- Schölkopf, B., & Smola, A. J. (2003). *Learning with kernel*. Cambridge: MIT Press.
- Schrodt, P. A. (2002). Mathematical modeling. In J. B. Manheim, R. C. Rich, & L. Willnat (Eds.), *Empirical political analysis: Research methods in political science* (5th ed.). New York: Longman.
- Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. . (1986). *Understanding regression analysis: An introductory guide*. Beverly Hills, CA: Sage.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.
- Schuman, H., & Presser, S. (1996). *Questions & answers in attitude surveys*. Thousand Oaks, CA: Sage.
- Schütz, A. (1963). Concept and theory formation in the social sciences. In M. A. Natanson (Ed.), *Philosophy of the social sciences* (pp. 231–249). New York: Random House.
- Schutz, A. (1964). Don Quixote and the problem of reality. In *Collected papers*, Vol. 2. The Hague: Martinus Nijhoff.
- Schütz, A. (1967). *Der sinnhafte Aufbau der sozialen Welt* [Phenomenology of the social world]. Evanston, IL: Northwestern University Press. (Originally published in 1932)
- Schutz, A. (1973). *Collected papers* (4 vols.). Dordrecht, The Netherlands: Kluwer.
- Schütz, A. (1976). Equality and the meaning structure of the social world. In *Collected Papers II* (pp. 226–273). The Hague: Martinus Nijhoff. (Originally published in 1955)
- Schütz, A. (1976). *The phenomenology of the social world* (G. Walsh & F. Lehnert, Trans.). London: Heinemann.
- Schwandt, T. (1996). Farewell to criteriology. *Qualitative Inquiry*, 2, 58–72.
- Schwandt, T. A. (1998). The interpretive review of educational matters: Is there any other kind? *Review of Educational Research*, 68(4), 409–412.
- Schwarz, N., & Sudman, S. (Eds.). (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. San Francisco: Jossey-Bass.
- Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, 49(3), 388–395.
- Scott, J. (1990). *A matter of record: Documentary sources in social research*. Cambridge, UK: Polity.
- Scott, J. (1992). *Social network analysis*. London: Sage.
- Scott, J. W. (1988). *Gender and the politics of history*. New York: Columbia University Press.
- Scott, M. B., & Lyman, S. (1968). Accounts. *American Sociological Review*, 33, 46–62.
- Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321–325.
- Scott, W. A. (1968). Attitude measurement. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology*. Reading, MA: Addison-Wesley.
- Scriven, M. (1959). Explanation and prediction in evolutionary theory. *Science*, 130, 477–482.
- Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.
- Scriven, M. (2001). Evaluation: Future tense. *The American Journal of Evaluation*, 22, 301–307.
- Seal, H. L. (1967). The historical development of the Gauss linear model. *Biometrika*, 54, 1–23.
- Seale, C. F. (1999). *The quality of qualitative research*. London: Sage.
- Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.
- Sechrest, L. (Ed.). (1979). *Unobtrusive measures today*. San Francisco: Jossey-Bass.
- Segrè, E. (1970). *Enrico Fermi: Physicist*. Chicago: University of Chicago Press.
- Sen, A. (1973). *On economic inequality*. New York: Norton.
- Sewell, W. H. (1942). The development of a sociometric scale. *Sociometry*, 5(3), 279–297.
- Shadish, W. R. (2000). The empirical program of quasi-experimentation. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy* (pp. 13–35). Thousand Oaks, CA: Sage.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
- Shadish, W. R., Jr., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.
- Shadish, W. R., Jr., Newman, D. L., Scheirer, M. A., & Wye, C. (1995). *Guiding principles for evaluators* (New Directions for Program Evaluation, No. 66). San Francisco: Jossey-Bass.
- Shapiro, G., & Markoff, J. (1998). *Revolutionary demands: A content analysis of the Cahier de Doléances of 1789*. Stanford, CA: Stanford University Press.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Shaw, C. (1966). *The jack roller*. Chicago: University of Chicago Press.
- Shaw, I., Bloor, M., Cormack, R., & Williamson, H. (1996). Estimating the prevalence of hard-to-reach populations: The illustration of mark-recapture methods in the study of homelessness. *Social Policy and Administration*, 30(1), 69–85.

- Sheskin, D. J. (1997). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: CRC Press.
- Shneiderman, B. (1992). Tree visualization with treemaps: A 2-D space-filling approach. *ACM Transactions on Graphics*, 11(1), 92–99.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.
- Shryock, H. S., Siegel, J. S., & Associates. (1973). *The methods and materials of demography*. New York: Academic Press. (Condensed edition by Edward G. Stockwell.)
- Si, M., Neufeld, R. R., & Dunbar, J. (1999). Removal of bedrails on a short-term nursing home rehabilitation unit. *The Gerontologist*, 39, 611–614.
- Sieber, J. E. (1992). *Planning ethically responsible research*. Newbury Park, CA: Sage.
- Sieber, J. E. (1996). Typically unexamined communication processes in research. In B. H. Stanley, J. E. Sieber, & G. B. Melton (Eds.), *Research ethics: A psychological approach*. Lincoln: University of Nebraska Press.
- Sieber, J. E. (Ed.). (1991). *Sharing social science data: Advantages and challenges*. Newbury Park, CA: Sage.
- Siegel, J. S. (2002). *Applied demography*. San Diego, CA: Academic Press.
- Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
- Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
- Silverman, D. (1993). *Interpreting qualitative data: Methods for analysing talk, text, and interaction*. London: Sage.
- Simon, H. A. (1954). Spurious correlation: A causal interpretation. *Journal of the American Statistical Association*, 49, 467–479.
- Simon, H. A. (1957). *Models of man*. New York: Wiley.
- Simon, H. A. (1976). Discussion: Cognition and social behavior. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior*. Hillsdale, NJ: Erlbaum.
- Simons, H. (1996). The paradox of case study. *Cambridge Journal of Education*, 26(2), 225–240.
- Sims, C. A. (1972). Money income and causality. *American Economic Review*, 62, 540–552.
- Singer, E. (1978). Informed consent: Consequences for response rate and response quality in social surveys. *American Sociological Review*, 43, 144–162.
- Singer, E., & Frankel, M. R. (1982). Informed consent procedures in telephone interviews. *American Sociological Review*, 47, 416–427.
- Singer, E., vonThurn, D., & Miller, E. (1995). Confidentiality assurances and response. *Public Opinion Quarterly*, 59, 66–77.
- Singleton, A. (1999). Measuring international migration: A case study of European cross-national comparisons. In D. Dorling & S. Simpson (Eds.), *Statistics in society: The arithmetic of politics* (pp. 148–158). London: Arnold.
- Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.). (1989). *Analysis of complex surveys*. Chichester, UK: Wiley.
- Skłodowska, E. (1992). *Testimonio hispanoamericano: Historia, teoría, poética*. New York: Peter Lang.
- Skoudis, E. (2002). *Counter Hack: A step-by-step guide to computer attacks and effective defenses*. Upper Saddle, NJ: Prentice Hall.
- Skvoretz, J., & Faust, K. (1999). Logit models for affiliation networks. *Sociological Methodology*, 29, 253–280.
- Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, 15(9), 5–11.
- Smart, B. (2000). Postmodern social theory. In B. Turner (Ed.), *The Blackwell companion to social theory* (pp. 447–480). Oxford, UK: Blackwell.
- Smelser, N. (1997). *Problems of sociology*. Berkeley: University of California Press.
- Smith, B. H. (1997). *Belief and resistance: Dynamics of contemporary intellectual controversy*. Cambridge, MA: Harvard University Press.
- Smith, D. (1997). Comment on Heckman's "Truth and method: Feminist standpoint theory revisited." *Signs*, 22(21): 392–397.
- Smith, D. P. (1992). *Formal demography*. New York: Plenum.
- Smith, G. (1997, Winter). Do statistics test scores regress toward the mean? *Chance*, pp. 42–45.
- Smith, J. (1993). *After the demise of empiricism: The problem of judging social and educational inquiry*. Norwood, NJ: Ablex.
- Smith, J. K. (1989). *The nature of social and educational inquiry: Empiricism versus interpretation*. Norwood, NJ: Ablex.
- Smith, J. K., & Heshusius, L. (1986). Closing down the conversation: The end of the quantitative-qualitative debate among educational enquirers. *Educational Researcher*, 15, 4–12.
- Smith, J., & Deemer, D. (2000). The problem of criteria in the age of relativism. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 877–922). Thousand Oaks, CA: Sage.
- Smith, L. M. (1978). An evolving logic of participant observation, educational ethnography and other case studies. *Review of Research in Education*, 6, 316–377.
- Smith, M. (1998). *Social science in question*. London: Sage.
- Smith, N. W. (2001). *Current systems in psychology*. Belmont, CA: Wadsworth.
- Smith, S. K., Tayman, J., & Swanson, D. A. (2001). *State and local population projections: Methodology and analysis*. New York: Kluwer Academic/Plenum Publishers.
- Smith, S., & Watson, J. (2001). *Reading autobiography: A guide for interpreting life narratives*. Minneapolis: University of Minnesota Press.

- Smith, T. W. (1992). Thoughts on the nature of context effects. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 163–184). New York: Springer-Verlag.
- Smithson, M. (1987). *Fuzzy set analysis for behavioral and social sciences*. New York: Springer-Verlag.
- Smithson, M. (2003). *Confidence intervals* (Sage University Papers on Quantitative Applications in the Social Sciences, 07–140). Thousand Oaks, CA: Sage.
- Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification*. San Francisco: Freeman.
- Snedecor, G. W. (1937). *Statistical methods*. Ames: Iowa State University Press.
- Snedecor, G. W. (1956). *Statistical methods* (5th ed.). Ames: Iowa State College Press.
- Snider, J. G., & Osgood, C. E. (Eds.). (1969). *Semantic differential technique: A sourcebook*. Chicago: Aldine.
- Snijders, T. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, 31, 361–395.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
- Snow, D. (1980). The disengagement process: A neglected problem in participant observation research. *Qualitative Sociology*, 3, 100–122.
- Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 1–38). New York: Plenum.
- Sobel, M., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), *Sociological methodology 1985* (pp. 152–178). San Francisco: Jossey-Bass.
- Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409–1438.
- Sokal, R. R., & Rohlf, F. J. (1981). *Biometry* (2nd ed.). San Francisco: W.H. Freeman.
- Sokal, R., & Sneath, P. (1963). *Principles of taxonomy*. San Francisco: Freeman.
- Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, 46, 137–150.
- Solomon, W. D. (1978). Ethics: Rules and principles. In W. T. Reich (Ed.), *Encyclopedia of bioethics*. New York: Free Press.
- Sørensen, A. B. (1979). A model and a metric for the analysis of the intragenerational status attainment process. *American Journal of Sociology*, 85, 361–384.
- Sørensen, J. B. (2002). The use and misuse of the coefficient of variation in organizational demography research. *Sociological Methods & Research*, 30, 475–491.
- Soulliere, D., Britt, D. W., & Maines, D. R. (2001). Conceptual modeling as a toolbox for grounded theorists. *Sociological Quarterly*, 42, 233–251.
- Soyland, A. J. (1994). *Psychology as metaphor*. London: Sage.
- Spanos, A. (1999). *Probability and statistical inference: Econometric modeling with observational data*. Cambridge, UK: Cambridge University Press.
- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15, 366–374.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 12, 271–295.
- Special section: Speaking in the name of the real: Freeman and Mead on Samoa. (1983). *American Anthropologist*, 85, 908–947.
- Spector, P. E. (1987). Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem? *Journal of Applied Psychology*, 72, 438–443.
- Spector, P. E. (1992). *Summated rating scale construction: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–082). Newbury Park, CA: Sage.
- Spector, P. E. (1994). Using self-report questionnaires in OB research: A comment on the use of a controversial method. *Journal of Organizational Behavior*, 15, 385–392.
- Spector, P. E., & Brannick, M. T. (1995). The nature and effects of method variance in organizational research. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology: 1995* (pp. 249–274). West Sussex, England: Wiley.
- Spector, P. E., Van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management*, 23, 659–677.
- Spiegel, M. R., Schiller, J., & Srinivasan, R. A. (2000). *Schaum's outline of probability and statistics* (2nd ed.). New York: McGraw-Hill.
- Spradley, J. P. (1979). *The ethnographic interview*. New York: Holt, Rinehart and Winston.
- SPSS base 10.0 user's guide. (1999). Chicago: SPSS.
- SPSS. (1990). *SPSS reference guide*. Chicago: SPSS.
- SPSS. (2001). *Reliability* (Technical section). Retrieved from www.spss.com/tech/stat/algorithms/11.0/reliability.pdf
- SPSS. (2002, August). Crosstabs. In *SPSS 11.0 statistical algorithms* [Online]. Available: <http://www.spss.com/tech/stat/algorithms/11.0/crosstabs.pdf>
- Stanley, L. (1992). *The auto/biographical I: Theory and practice of feminist auto/biography*. Manchester, UK: Manchester University Press.
- Stanley, L., & Wise, S. (1993). *Breaking out again: Feminist ontology and epistemology*. London: Routledge Kegan Paul.
- Stanton, A. L., Burkner, E. J., & Kershaw, D. (1991). Effects of researcher follow-up of distressed subjects: Tradeoff between validity and ethical responsibility? *Ethics & Behavior*, 1(2), 105–112.
- Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

- Stem, D. E., Jr., & Lamb, C. W., Jr. (1981). The marble-drop technique: A procedure for gathering sensitive information. *Decision Sciences*, 12, 702–707.
- Stern, J., Stackowiack, R., & Greenwald, R. (2001). *Oracle essentials: Oracle9i, Oracle8i and Oracle 8*. Sebastopol, CA: O'Reilly.
- Stevens, J. (2001). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.
- Stevens, S. S. (1951). *Handbook of experimental psychology*. New York: Wiley.
- Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science*, 161, 849–861.
- Stewart, D. K., & Love, W. A. (1968). A general canonical correlation index. *Psychological Bulletin*, 70, 160–163.
- Stewart, D. W., & Shamdasani, P. N. (1990). *Focus groups: Theory and practice* (Applied Social Research Methods Series, Vol. 20). Newbury Park, CA: Sage.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press of Harvard University Press.
- Stigler, S. (1986). Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1, 359–378.
- Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.
- Stoll, D. (1999). *Rigoberta Menchú and the story of all poor Guatemalans*. Boulder, CO: Westview.
- Stone, A. A., Shiffman, S., & DeVries, M. (1999). Ecological momentary assessment. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 27–38). New York: Russell Sage.
- Stone, A. A., Turkkan, J. S., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (Eds.). (2000). *The science of self-report: Implications for research and practice*. Mahwah, NJ: Lawrence Erlbaum.
- Stouffer, S. A. (1950). Some observations on study design. *American Journal of Sociology*, 55, 355–361.
- Strauss, A. (1987). *Qualitative analysis for social scientists*. New York: Cambridge University Press.
- Strauss, A. L., Corbin, J., Fagerhaugh, S., Glaser, B., Maines, D., Suczek, B., & Weiner, C. (1984). *Chronic illness and the quality of life*. St. Louis, MO: Mosby.
- Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.
- Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Grounded theory procedures and techniques* (2nd ed.). Thousand Oaks, CA: Sage.
- Stringer, E. T. (1996). *Action Research: A handbook for practitioners*. Thousand Oaks, CA: Sage.
- Strohmetz, D. B., & Rosnow, R. L. (1994). A mediational model of research artifacts. In J. Brzezinski (Ed.), *Probability in theory-building: Experimental and nonexperimental approaches to scientific research in psychology* (pp. 177–196). Amsterdam: Editions Rodopi.
- Stuart, A. (1984). *The ideas of sampling*. London: Griffin.
- Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics* (Vol. 1). London: Edward Arnold.
- Stuart, A., Ord, J. K., & Arnold, S. (1999). *Kendall's advanced theory of statistics: Vol. 2A. Classical inference & the linear model* (6th ed.). New York: Oxford University Press.
- Studenmund, A. H. (1992). *Using econometrics: A practical guide* (2nd ed.). New York: HarperCollins.
- Student. (1908). The probable error of a mean. In E. S. Pearson & J. Wishart (Eds.), *Student's collected papers*. London: University College.
- Suchman, E. A. (1967). *Evaluative research: Principles and practice in public service and social action programs*. New York: Russell Sage.
- Sudman, S. (1976). *Applied sampling*. New York: Academic Press.
- Sudman, S., & Bradburn, N. (1982). *Asking questions*. San Francisco: Jossey-Bass.
- Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.
- Sudnow, D. (1978). *Ways of the hand*. Cambridge, MA: Harvard University Press.
- Sudnow, D. (2001). *Ways of the hand* (Rev. ed.). Cambridge, MA: MIT Press.
- Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioural observation data*. London: Lawrence Erlbaum.
- Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North Holland.
- Swokowski, E. W. (1979). *Calculus with analytic geometry* (2nd ed.). Boston: Prindle, Weber & Schmidt.
- Symon, G. J., & Clegg, C. W. (1991). Technology-led change: A study of the implementation of CAD/CAM. *Journal of Occupational Psychology*, 64, 273–290.
- Szalai, A. (Ed.). (1972). *The use of time*. The Hague: Mouton.
- Sztompka, P. (1990). *Robert Merton: An intellectual profile*. London: Macmillan.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Computer-assisted research design and analysis*. Boston: Allyn and Bacon.
- Tacq, J. J. A. (1984). *Causaliteit in Sociologisch Onderzoek. Een Beoordeling van Causale Analysetechnieken in het Licht van Wijsgerige Opvattingen over Causaliteit* [Causality in sociological research: An evaluation of causal techniques of analysis in the light of philosophical theories of causality]. Deventer, The Netherlands: Van Loghum Slaterus.
- Tacq, J. J. A. (1997). *Multivariate analysis techniques in social science research: From problem to analysis*. London: Sage.
- Takane, Y. (1981). Multidimensional scaling of sorting data. In Y. P. Chaubey & T. D. Dwivedi (Eds.), *Topics in applied statistics*. Montreal: Concordia University Press.
- Takane, Y., Young, F. W., & DeLeeuw, J. (1977). Non-metric individual differences multidimensional scaling: An

- alternating least squares method with optimal scaling features. *Psychometrika*, 42, 7–67.
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528–550.
- Taris, T. W. (2000). *A primer in longitudinal data analysis*. London: Sage.
- Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.
- Taylor, C. (1994). The politics of recognition. In C. Taylor & A. Gutmann (Eds.), *Multiculturalism and the politics of recognition* (pp. 25–73). Princeton, NJ: Princeton University Press.
- Taylor, M. F., with Brice, J., Buck, N., & Prentice-Lane, E. (Eds.). (2000). *British Household Panel Survey user manual volume B9: Codebook*. Colchester, UK: University of Essex.
- Taylor, S. J., & Bogdan, R. (1998). *Introduction to qualitative research methods: A guide & resource*. New York: Wiley.
- Tedlock, D. (1999). Poetry and ethnography: A dialogical approach. *Anthropology and Humanism*, 24(2), 155–167.
- ten Have, P. (1999). *Doing conversation analysis*. Thousand Oaks, CA: Sage.
- Tenenhaus, M. (1998). *La régression PLS* [PLS regression]. Paris: Technip.
- The 1980 U.S. census [Special issue]. (1992). *Survey Methodology*, 18.
- The 1990 U.S. census [Special issue]. (1993). *Journal of the American Statistical Association*, 88.
- The 1990 U.S. census [Special issue]. (1994). *Statistical Science*, 9.
- The 2000 U.S. census. (2001). *Society*, 39, 2–53.
- Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. New York: Springer-Verlag.
- Theus, M. (1997). Visualization of categorical data. In W. Bandilla & F. Faulbaum (Eds.), *SoftStat '97: Advances in statistical software* (Vol. 6, pp. 47–55). Stuttgart, Germany: Lucius & Lucius.
- Thom, R. (1975). *Structural stability and morphogenesis*. Reading, MA: Benjamin.
- Thomas, H. (1997). Dancing: Representation and difference. In J. McGuigan (Ed.), *Cultural methodologies*. London: Sage.
- Thomas, J. (1992). *Doing critical ethnography*. Newbury Park, CA: Sage.
- Thomas, W. I., & Znaniecki, F. (1918–1920). *The Polish peasant in Europe and America* (5 vols.). New York: Dover.
- Thomas, W. I., & Znaniecki, F. (1958). *The Polish peasant in Europe and America* (2 vols.). New York: Dover. (Original five-volume work published in 1918–1920).
- Thompson, B. (1985). Alternate methods for analyzing data from experiments. *Journal of Experimental Education*, 54, 50–55.
- Thompson, B. (1997). The importance of structure coefficients in structural equation modeling confirmatory factor analysis. *Educational and Psychological Measurement*, 57, 5–19.
- Thompson, B. (2000). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 285–316). Washington, DC: American Psychological Association.
- Thompson, J. (1981). *Critical hermeneutics: A study in the thought of Paul Ricoeur and Jürgen Habermas*. Cambridge, UK: Cambridge University Press.
- Thompson, J. (1990). *Ideology and modern culture*. Stanford, CA: Stanford University Press.
- Thompson, P. (2000). *The voice of the past* (3rd ed.). Oxford, UK: Oxford University Press.
- Thorndike, R. L. (1967). The analysis and selection of test items. In D. Jackson & S. Messick (Eds.), *Problems in human assessment* (pp. 201–216). New York: McGraw-Hill.
- Thorne, S. (1994). Secondary analysis in qualitative research: Issues and implications. In J. Morse (Ed.), *Critical issues in qualitative research methods* (pp. 263–279). Thousand Oaks, CA: Sage.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433–451.
- Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology*, 17, 446–457.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273–286.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 23, 529–554.
- Thurstone, L. L. (1929). Fechner's law and the method of equal-appearing intervals. *Journal of Experimental Psychology*, 12, 214–224.
- Thurstone, L. L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology*, 26, 249–269.
- Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
- Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitudes*. Chicago: University of Chicago Press.
- Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology*. Monterey, CA: Brooks/Cole.
- Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, 20, 410–433.
- Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22, 358–376.
- Tintner, G. (1968). *Methodology and mathematical economics and econometrics*. Chicago: University of Chicago Press.
- Titon, J. (1980). The life story. *Journal of American Folklore*, 93(369), 276–292.
- Titscher, S., Meyer, M., Wodak, R., & Vetter, R. (2000). *Methods of text and discourse analysis*. London: Sage.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24–36.
- Tolley, H. D., & Manton, K. G. (1992). Large sample properties of estimates of discrete grade of membership model. *Annals of the Institute of Statistical Mathematics*, 41, 85–95.
- Tolley, H. D., Kovtun, M., Manton, K. G., & Akushevich, I. (2003). Statistical properties of grade of membership

- models for categorical data. Manuscript to be submitted to *Proceedings of National Academy of Sciences*, 2003.
- Tommerup, P. (1993). *Adhocratic traditions, experience narratives and personal transformation: An ethnographic study of the organizational culture and folklore of the Evergreen State College, an innovative liberal arts college*. Unpublished doctoral dissertation, University of California, Los Angeles.
- Toombs, S. K. (2001). Reflections on bodily change: The lived experience of disability. In S. K. Toombs (Ed.), *Handbook of phenomenology and medicine* (pp. 247–261). Boston: Kluwer Academic.
- Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park, CA: Sage.
- Toothaker, L. E. (1993). *Multiple comparison procedures*. Newbury Park, CA: Sage.
- Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.
- Toulmin, S. E. (1978). Science, philosophy of. In *Encyclopaedia Britannica* (15th ed., Vol. 16, pp. 375–393). Chicago: Britannica.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.
- Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin*, 96, 394–401.
- Townsend, P., Davidson, N., & Whitehead, M. (1988). *Inequalities in health: The Black report and the health divide*. Harmondsworth: Penguin.
- Tracy, K. (2002). *Everyday talk: Building and reflecting identities*. New York: Guilford.
- Tracy, P. E., & Fox, J. A. (1981). The validity of sensitive measurements: A comparison of two measurement strategies. *American Sociological Review*, 46, 187–200.
- Train, K. (1986). *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*. Cambridge: MIT Press.
- Tranfield, D., & Starkey, K. (1998). The nature, organization and promotion of management research: Towards policy. *British Journal of Management*, 9, 341–353.
- Traugott, M. W., & Katosh, J. P. (1979). Response validity in surveys of voting behavior. *Public Opinion Quarterly*, 43(3), 359–377.
- Trice, A. D. (1987). Informed consent: VII: Biasing of sensitive self-report data by both consent and information. *Journal of Social Behavior and Personality*, 2, 369–374.
- Triplitt, N. (1898). The dynamogenic factors in pacemaking and competition. *American Journal of Psychology*, 9, 507–533.
- Troyna, B. (1994). Reforms, research and being reflexive about being reflective. In D. Halpin & B. Troyna (Eds.), *Researching education policy*. London: Falmer.
- Trussell, J., & Rodríguez, G. (1990). Heterogeneity in demographic research. In J. Adams, D. A. Lam, A. I. Hermalin, & P. E. Smouse (Eds.), *Convergent issues in genetics and demography* (pp. 111–132). Oxford, UK: Oxford University Press.
- Truzzi, M. (1974). *Verstehen: Subjective understanding in the social sciences*. Reading, MA: Addison-Wesley.
- Tsay, R. S. (2002). *Analysis of financial time series*. New York: Wiley.
- Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72, 20–22.
- Tsuji, R. (1999). Trusting behavior in prisoner's dilemma: The effects of social networks. *Dissertation Abstracts International Section A: Humanities & Social Sciences*, 60(5-A), 1774. (UMI No. 9929113)
- Tuchman, G. (1994). Historical social science: Methodologies, methods, and meanings. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 306–323). Thousand Oaks, CA: Sage.
- Tucker, L. R. (1960). Intra-individual and inter-individual multidimensionality. In H. Gulliksen & S. Messick (Eds.), *Psychometric scaling: Theory and applications* (pp. 155–167). New York: Wiley.
- Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
- Tukey, J. W. (1958). Bias and confidence in not quite large samples [abstract]. *Annals of Mathematical Statistics*, 29, 614.
- Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Tuma, N. B., & Hannan, M. T. (1984). *Social dynamics: Models and methods*. Orlando, FL: Academic Press.
- Turner, A. C. (1982). What subjects of survey research believe about confidentiality. In J. E. Sieber (Ed.), *The ethics of social research: Surveys and experiments* (pp. 151–165). New York: Springer-Verlag.
- Turner, R. H. (1951). The quest for universals. *American Sociological Review*, 18(6), 604–611.
- Turner, S. P., & Factor, R. A. (1981). Objective possibility and adequate causality in Weber's methodological writings. *Sociological Review*, 29(1), 5–28.
- Turner, V., & Bruner, E. (Eds.). (1986). *The anthropology of experience*. Urbana: University of Illinois Press.
- Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, 59, 251–278.
- Tzelgov, J., & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin*, 109, 524–536.
- U.S. Census Bureau. (2001). *Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy*

- on Adjustment for Non-Redistricting Uses (With supporting documentation, Reps. 1–24). Washington, DC: Government Printing Office.
- U.S. Federal Regulations of Human Research 45 CFR 46. Retrieved from <http://ohrp.osophy.dhhs.gov>.
- UK Data Archive (n.d.). [Online]. Available: <http://www.data-archive.ac.uk/>.
- United Nations Statistics Division. (2001). *Handbook on census management for population and housing censuses—Series F* (No. 83, Revision 1). New York: United Nations.
- United Nations. (1995). *World population prospects: The 1994 revision*. New York: Author.
- United Nations. (1996). *Demographic yearbook 1994*. New York: Author.
- University of Mississippi. (n.d.). PsychExperiments. Retrieved from <http://psychexps.olemiss.edu>
- Utts, J. M. (1996). *Seeing through statistics*. Belmont, CA: Duxbury.
- Valero, P., Young, F., & Friendly, M. (2003). Visual categorical analysis in ViSta. *Computational Statistics and Data Analysis*, 43(4), 495–508.
- Vallet, L.-A. (2001). Forty years of social mobility in France. *Revue Française de Sociologie: An Annual English Selection*, 42(Suppl.), 5–64.
- Valsiner, J., & Benigni, L. (1986). Naturalistic research and ecological thinking in the study of child development. *Developmental Review*, 6(3), 203–223.
- Van de Geer, J. (1971). *Introduction to multivariate analysis for the social sciences*. San Francisco: Freeman.
- Van den Ende, H. W. (1971). *Beschrijvende Statistiek voor Gedragwetenschappen*. Amsterdam/Brussel: Agon Elsevier.
- van der Ark, L. A., & van der Heijden, P. G. M. (1998). Graphical display of latent budget analysis and latent class analysis, with special reference to correspondence analysis. In M. Greenacre & J. Blasius (Eds.), *Visualization of categorical data* (pp. 489–508). San Diego: Academic Press.
- van der Ark, L. A., van der Heijden, P. G. M., & Sikkil, D. (1999). On the identifiability of the latent model. *Journal of Classification*, 16, 117–137.
- van der Heijden, P. G. M., Mooijaart, A., & de Leeuw, J. (1992). Constrained latent budget analysis. In P. Marsden (Ed.), *Sociological methodology* (pp. 279–320). Cambridge, UK: Basil Blackwell.
- van der Heijden, P. G. M., van der Ark, L. A., & Mooijaart, A. (2002). Some examples of latent budget analysis and its extensions. In J. A. Hagenaars & A. McCutcheon (Eds.), *Applied latent class analysis* (pp. 107–136). Cambridge, UK: Cambridge University Press.
- Van der Heijden, P. G. M., van Gils, G., Bouts, J., & Hox, J. J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods & Research*, 28, 505–537.
- Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.
- Van Dijk, T. A. (Ed.). (1985). *Handbook of discourse analysis: Vols. 1–4*. London: Academic Press.
- Van Maanen, J. (1979). The fact and fiction in organizational ethnography. *Administrative Science Quarterly*, 24, 539–550.
- Van Maanen, J. (1988). *Tales of the field: On writing ethnography*. Chicago: University of Chicago Press.
- van Manen, M. (1997). *Researching lived experience: Human science for an action sensitive pedagogy*. London, Ontario: Althouse.
- van Teijlingen, E., & Hundley, V. (2002). The importance of pilot studies. *Nursing Standard*, 16(40), 33–36.
- Vapnik, V. N. (1999). *Statistical learning theory*. New York: Wiley.
- Vaupel, J. W. (1990). Relatives' risks: Frailty models of life history data. *Theoretical Population Biology*, 37, 220–234.
- Vaupel, J. W., & Yashin, A. I. (1985). The deviant dynamics of death in heterogenous populations. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 179–211). San Francisco: Jossey-Bass.
- Vaupel, J. W., & Yashin, A. I. (1985). Heterogeneity's ruses: Some surprising effects of selection on population dynamics. *American Statistician*, 39, 176–185.
- Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439–454.
- Velleman, P. F. (1988). *DataDesk version 6.0 statistics guide*. Ithaca, NY: Data Description, Inc.
- Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *American Statistician*, 47, 65–72.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer-Verlag.
- Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.
- Verbrugge, L. (1980). Health diaries. *Medical Care*, 18, 73–95.
- Vermunt, J. K. (1997). *Log-linear models for event histories*. Thousand Oaks, CA: Sage.
- Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge, UK: Cambridge University Press.
- Vermunt, J. K., Rodrigo, M. F., & Ato-Garcia, M. (2001). Modeling joint and marginal distributions in the analysis of categorical panel data. *Sociological Methods & Research*, 30, 170–196.
- Vogt, P. W. (1993). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences*. Newbury Park, CA: Sage.
- Vogt, W. P. (1999). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (2nd ed.). Thousand Oaks, CA: Sage.

- Von Collani, E., & Drager, K. (2001). *Binomial distribution handbook*. Basel, Switzerland: Birkhauser.
- von Glasersfeld, E. (1984). An introduction to radical constructivism. In P. Watzlawick (Ed.), *The invented reality*. New York: Norton.
- von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behaviour*. Princeton, NJ: Princeton University Press.
- von Wright, G. H. (1971). *Explanation and understanding*. Ithaca, NY: Cornell University Press.
- Wachter, K. W., & Trussell, J. (1982). Estimating historical heights. *Journal of the American Statistical Association*, 77(378), 279–293.
- Wackerly, D. D., Mendelhall, W., III, & Scheaffer, R. L. (1996). *Mathematical statistics with applications* (5th ed.). Belmont, CA: Duxbury.
- Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.
- Walby, S., & Olsen, W. (2002, November). *The impact of women's position in the labour market on pay and implications for UK productivity*. Cabinet Office Women and Equality Unit. Retrieved from www.womenandequalityunit.gov.uk.
- Wald, A., & Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Annals of Mathematical Statistics*, 11, 147–162.
- Walker, E., & Lev, J. (1953). *Statistical inference*. New York: Henry Holt.
- Walker, R. (Ed.). (1985). *Applied qualitative research*. Aldershot, UK: Gower.
- Walkerline, V., Lucey, H., & Melody, J. (2001). *Growing up girl: Psycho-social explorations of gender and class*. Basingstoke, UK: Palgrave.
- Wallgren, A. et al. (1996). *Graphing statistics and data: Creating better data*. Thousand Oaks, CA: Sage.
- Walsh, A., & Ollenburger, J. C. (2001). *Essential statistics for the social and behavioral sciences: A conceptual approach*. Upper Saddle River, NJ: Prentice Hall.
- Walt, S. M. (1987). *The origins of alliances*. Ithaca, NY: Cornell University Press.
- Walton Braver, M., & Braver, S. L. (1996). Statistical treatment of the Solomon four-group design: A meta-analytic approach. *Psychological Bulletin*, 104(1), 150–154.
- Walton, J. (1992). Making the theoretical case. In C. C. Ragin & H. S. Becker (Eds.), *What is a case? Exploring the foundations of social inquiry* (pp. 121–137). New York: Cambridge University Press.
- Ward, A. (2002). *Oral history and copyright*. Retrieved June 7, 2002 from <http://www.oralhistory.org.uk>
- Warde, A., Martens, L., & Olsen, W. K. (1999). Consumption and the problem of variety: Cultural omnivorousness, social distinction and dining out. *Sociology*, 30(1), 105–128.
- Warner, R. M. (1998). *Spectral analysis of time-series data*. New York: Guilford.
- Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63–69.
- Warren, C. A. B. (1987). *Madwives: Schizophrenic women in the 1950s*. New Brunswick, NJ: Rutgers University Press.
- Warren, C. A. B. (2002). Qualitative interviewing. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research: Context and method* (pp. 83–101). Thousand Oaks, CA: Sage.
- Warren, C. A. B., & Hackney, J. K. (2000). *Gender issues in ethnography* (Qualitative Research Methods Series vol. 9, 2nd ed.). Thousand Oaks, CA: Sage.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.
- Wasserman, S., & Galaskiewicz, J. (Eds.). (1994). *Advances in social network analysis*. Thousand Oaks, CA: Sage.
- Watson, J. B. (1959). *Behaviorism*. Chicago: University of Chicago Press.
- Watson, M. W. (1994). Vector autoregression and cointegration. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of econometrics* (Vol. 4, pp. 2843–2915). Amsterdam: Elsevier.
- Watts, D. (1999). *Small worlds: The dynamics of networks between order and randomness*. Princeton, NJ: Princeton University Press.
- Weaver, L., & Cousins, B. (2001, November). *Unpacking the participatory process*. Paper presented at the annual meeting of the American Evaluation Association, St. Louis, MO.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive measures in the social sciences*. Chicago: Rand McNally.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. B. (1981). *Nonreactive measures in the social sciences* (2nd ed.). Boston: Houghton Mifflin.
- Webb, J. F., Khazen, R. S., Hanley, W. B., Partington, M. S., Percy, W. J. L., & Rathborn, J. C. (1973). PKU screening—is it worth it? *Canadian Medical Association Journal*, 108, 328–329.
- Weber, M. (1949). Die “Objektivität” sozialwissenschaftlicher und sozialpolitischer Erkenntnis [“Objectivity” in social science and social policy]. In *The methodology of the social sciences: Max Weber* (pp. 50–112). New York: Free Press. (Originally published in 1904)
- Weber, M. (1964). *The theory of social and economic organization* (A. M. Henderson & T. Parsons, Trans.). New York: Free Press.
- Weber, M. (1975). *Roscher and Knies*. New York: Free Press.
- Weber, R. P. (1990). *Basic content analysis*. Newbury Park, CA: Sage.
- Wedel, M., & DeSarbo, W. S. (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 352–388). Cambridge, UK: Basil Blackwell.

- Weisberg, H. F. (1974). Dimensionland: An excursion into spaces. *American Journal of Political Science*, 18, 743–776.
- Weisberg, H. F. (1974). Models of statistical relationship. *American Political Science Review*, 68, 1638–1655.
- Weisberg, H. F. (1992). *Central tendency and variability* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–083). Newbury Park, CA: Sage.
- Weisberg, S. (1985). *Applied linear regression*. New York: Wiley.
- Weiss, C. (1977). *Use of social research in public policy*. Lexington, MA: D. C. Heath.
- Weiss, C. H. (1972). A treeful of owls. In C. H. Weiss (Ed.), *Evaluating action programs* (pp. 3–27). Boston: Allyn & Bacon.
- Weiss, C. H., & Bucuvalas, M. (1980). Truth test and utility test: Decision makers' frame of reference for social science research. *American Sociological Review*, 45, 302–313.
- Weiss, H. M. (2002). Deconstructing job satisfaction: Separating evaluations, beliefs and affective experiences. *Human Resources Management Review*, 12, 173–194.
- Weiss, R. (1995). *Learning from strangers*. New York: Free Press.
- Weitzman, E. A., & Miles, M. B. (1995). *Computer programs for qualitative data analysis: A software source book*. Thousand Oaks, CA: Sage.
- Welch, B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika*, 34, 28–35.
- Weller, S. C., & Romney, A. K. (1988). *Systematic data collection* (Qualitative Research Methods, Vol. 10). Newbury Park, CA: Sage.
- Weller, S. C., & Romney, A. K. (1990). *Metric scaling: Correspondence analysis*. Newbury Park, CA: Sage.
- Wellman, B., & Berkowitz, S. D. (Eds.). (1997). *Social structures: A network approach* (updated ed.). Greenwich, CT: JAI.
- Wengraf, T. (2001). *Qualitative research interviewing: Biographic-narrative and semi-structured method*. London: Sage.
- Werner, O., & Schoepfle, G. M. (1987). *Systematic fieldwork* (2 vols.). Newbury Park, CA: Sage.
- West, S. G., Hepworth, J. T., McCall, M. A., & Reich, J. W. (1989). An evaluation of Arizona's July 1982 drunk driving law: Effects on the City of Phoenix. *Journal of Applied Social Psychology*, 19, 1212–1237.
- Wetherell, M., & Potter, J. (1992). *Mapping the language of racism: Discourse and the legitimization of exploitation*. Brighton, UK: Harvester Wheatsheaf.
- Wetherell, M., Taylor, S., & Yates, S. J. (Eds.). (2001). *Discourse theory and practice*. London: Sage.
- Wheldall, K., & Merrett, F. (1989). *Positive teaching: The behavioural approach*. Birmingham, UK: Positive Products.
- Wherry, R. J., Sr. (1984). *Contributions to correlational analysis*. New York: Academic Press.
- Whewell, W. (1847). *The philosophy of the inductive sciences* (Vols. 1 & 2). London: Parker.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817–838.
- White, H. (1984). *Asymptotic theory for econometricians*. Orlando, FL: Academic Press.
- White, L. (1999). *Political analysis: Technique and practice* (4th ed.). Orlando, FL: Harcourt Brace.
- Whiting, J. W., Child, I. L., & Lambert, W. W. (1966). *Field guide for the study of socialization*. New York: Wiley.
- Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Chichester, UK: Wiley.
- Whyte, W. F. (1955). *Street corner society: The social structure of an Italian slum* (2nd ed.). Chicago: University of Chicago Press. (Original work published 1943)
- Whyte, W. F. (1984). *Learning from the field: A guide from experience*. Beverly Hills, CA: Sage.
- Whyte, W. F. (Ed.). (1991). *Participatory action research*. Newbury Park, CA: Sage.
- Wiggershaus, R. (1995). *The Frankfurt School: Its history, theories and political significance* (M. Robertson, Trans.). Cambridge, UK: Polity.
- Wiggins, L. M. (1973). *Panel analysis*. Amsterdam: Elsevier.
- Wilcox, R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York: Springer.
- Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.
- Wilcox, R. R. (2001). *Fundamentals of modern statistical methods*. New York: Springer.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80–83.
- Wildt, A. R., & Ahtola, O. T. (1978). *Analysis of covariance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–012). Beverly Hills, CA: Sage.
- Waller, D. (1967). *Scientific sociology: Theory and method*. Englewood Cliffs, NJ: Prentice Hall.
- Williams, G. (1984). The genesis of chronic illness: Narrative re-construction. *Sociology of Health & Illness*, 6(2), 175–200.
- Williams, L. J., & Brown, B. K. (1994). Method variance in organizational behavior and human resources research: Effects on correlations, path coefficients, and hypothesis testing. *Organizational Behavior and Human Decision Processes*, 57, 185–209.
- Williams, M. (1998). The social world as knowable. In T. May & M. (Eds.), *Knowing the social world* (pp. 5–21). Buckingham, UK: Open University Press.
- Williams, M. (2000). Interpretivism and generalisation. *Sociology*, 34(2), 209–224.
- Williams, M. (2000). *Science and social science: An introduction*. London: Routledge.
- Williams, M., & May, T. (1996). *Introduction to the philosophy of social research*. London: UCL Press.

- Williams, R. (1977). *Marxism and literature*. Oxford, UK: Oxford University Press.
- Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal*, 19, 249–258.
- Wilson, P. J. (1974). *Oscar: An inquiry into the nature of sanity*. New York: Vintage.
- Wilson, T. P. (1969). A proportional reduction in error interpretation for Kendall's tau-b. *Social Forces*, 47, 340–342.
- Winch, P. (1990). *The idea of a social science and its relation to philosophy*. London: Routledge. (Originally published in 1958)
- Winer, B. J. (1962). *Statistical principles in experimental designs*. New York: McGraw-Hill.
- Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.
- Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology*, 18, 327–350.
- Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 659–706.
- Winship, C., & Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research*, 23(2), 230–257.
- Wiseman, F., Moriarty, M., & Schafer, M. (1975–1976). Estimating public opinion with the randomized response model. *Public Opinion Quarterly*, 39(4), 507–513.
- Wishart, J. (1938). Growth-rate determinations in nutritional studies with the Bacon pig, and their analysis. *Biometrika*, 30, 16–28.
- Wittgenstein, L. (1974). *On certainty* (G. E. M. Anscombe & G. H. Von Wright, Eds.). Oxford, UK: Blackwell.
- Wodak, R., & Meyer, M. (2001). *Methods in critical discourse analysis*. Thousand Oaks, CA: Sage.
- Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate analysis* (pp. 391–420). New York: Academic Press.
- Wolf, M. (1992). *A thrice told tale: Feminism, postmodernism and ethnographic responsibility*. Stanford, CA: Stanford University Press.
- Wolfe, J. H. (1970). Pattern clustering by multivariate mixtures. *Multivariate Behavioral Research*, 5, 329–350.
- Wonnacott, T. H., & Wonnacott, R. J. (1990). *Introductory statistics for business and economics* (4th ed.). San Francisco: Jossey-Bass.
- Wood, B. D. (1992). Modeling federal implementation as a system: The clean air case. *American Journal of Political Science*, 1, 40–67.
- Wood, J. W., & Weinstein, M. (1990). Heterogeneity in fecundability: The effect of fetal loss. In J. Adams, D. A. Lam, A. I. Hermalin, & P. E. Smouse (Eds.), *Convergent issues in genetics and demography* (pp. 171–188). Oxford, UK: Oxford University Press.
- Wood, J. W., Holman, D. J., Yasin, A., Peterson, R. J., Weinstein, M., & Chang, M.-C. (1994). A multistate model of fecundability and sterility. *Demography*, 31, 403–426.
- Wood, S. (2001). mgcv: GAMs and generalized ridge regression for R. *R News*, 1, 20–25.
- Woodbury, M. A., & Clive, J. (1974). Clinical pure types as a fuzzy partition. *Journal of Cybernetics*, 4(3), 111–121.
- Woodward, J. (1995). Causation and explanation in econometrics. In D. Little (Ed.), *On the reliability of economic models: Essays in the philosophy of economics* (pp. 9–61). Boston: Kluwer Academic.
- Wooldridge, J. M. (2000). *Introductory econometrics: A modern approach*. Cincinnati, OH: South-Western.
- Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.
- Wooldridge, J. M. (2003). *Introductory econometrics: A modern approach* (2nd ed.). Cincinnati, OH: South-Western College Publishing.
- Woolgar, S. (Ed.). (1988). *Knowledge and reflexivity*. London: Sage.
- World Bank, International Economics Department. (1997). *World tables of economic and social indicators, 1950–1992* [Computer file]. Ann Arbor, MI: Interuniversity Consortium for Political and Social Science Research.
- Wright, B. D. (1997). S. S. Stevens revisited. *Rasch Measurement Transactions*, 11, 552–553.
- Wright, S. (1918). On the nature of size factors. *Genetics*, 3, 367–374.
- Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, 5, 161–215.
- Wright, W. (1975). *Six guns and society*. Berkeley: University of California Press.
- Wu, L. L., & Martinson, B. C. (1993). Family structure and the risk of a premarital birth. *American Sociological Review*, 58(2), 210–232.
- Wu, L. L., & Tuma, N. B. (1990). Local hazard models. In C. C. Clogg (Ed.), *Sociological methodology 1990* (pp. 141–180). Oxford, UK: Basil Blackwell.
- Wylie, R. (1974). *The self concept: A review of methodological considerations and measuring instruments*. Lincoln: University of Nebraska Press.
- Xie, Y. (1989). An alternative purging method: Controlling the composition-dependent interaction in an analysis of rates. *Demography*, 26, 711–716.
- Xie, Y. (1992). The log-multiplicative layer effect model for comparing mobility tables. *American Sociological Review*, 57, 380–395.
- Xie, Y., & Pimentel, E. E. (1992). Age patterns of marital fertility: Revising the Coale-Trussell method. *Journal of the American Statistical Association*, 87, 977–984.
- Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association*, 29, 51–66.
- Yeo, I.-K., & Johnson, R. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87, 954–959.

- Yin, R. K. (1989). *Case study research: Design and methods*. Newbury Park, CA: Sage.
- Yin, R. K. (1994). *Case study research: Design and methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Yin, R. K. (1998). The abridged version of case study research. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 229–259). Thousand Oaks, CA: Sage.
- Young, F. W., & Hamer, R. M. (1987). *Multidimensional scaling: History, theory, and applications*. Hillsdale, NJ: Lawrence Erlbaum.
- Young, K., Ashby, D., Boaz, A., & Grayson, L. (2002). Social science and the evidence-based policy movement. *Social Policy and Society*, 1(3), 215–224.
- Yuen, K. K. (1974). The two-sample trimmed *t* for unequal populations variances. *Biometrika*, 61, 165–170.
- Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, 2, 121–134.
- Yule, G. U. (1971). On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. In A. Stuart & M. Kendall (Eds.), *Statistical papers of George Udny Yule* (pp. 389–420). New York: Hafner. (Original work published 1927)
- Yule, G. U., & Kendall, M. G. (1950). *An introduction to the theory of statistics* (14th ed.). London: Griffin.
- Yule, G. U., & Kendall, M. G. (1968). *An introduction to the theory of statistics*. New York: Hafner.
- Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8, 338–353.
- Zaller, J. R. (1992). *The nature and origins of mass opinion*. Cambridge, UK: Cambridge University Press.
- Zaman, A. (1996). *Statistical foundations for econometric techniques*. New York: Academic Press.
- Zanna, M. P., & Cooper, J. (1974). Dissonance and the pill: An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology*, 29, 703–709.
- Zeeman, E. C. (1972). Differential equations for the heartbeat and nerve impulse. In C. H. Waddington (Ed.), *Towards a theoretical biology* (Vol. 4, pp. 8–67). Edinburgh, UK: Edinburgh University Press.
- Zeisel, H., & Kaye, D. H. (1997). *Prove it with figures*. New York: Springer.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57, 348–368.
- Zellner, A., & Theil, H. (1962). Three-stage least squares: Simultaneous estimation of simultaneous equations. *Econometrica*, 30, 54–78.
- Zetenyi, T. (Ed.). (1988). *Fuzzy sets in psychology*. Amsterdam: North-Holland.
- Zimmerman, D. H., & Wieder, D. (1977). The diary-interview method. *Urban Life*, 5, 479–498.
- Zimmerman, M. (2001–2002). Rigoberta Menchú, David Stoll, subaltern narrative and testimonial truth: A personal testimony. *Antípodas: Journal of Hispanic and Galician Studies*, 13/14, 103–124.
- Znaniecki, F. (1934). *The method of sociology*. New York: Farrar and Rhienhart.
- Znaniecki, F. (1969). *On humanistic sociology: Selected papers* (R. Bierstadt, Ed.). Chicago: University of Chicago Press.
- Zorn, C. J. W. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, 45, 470–490.
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432–442.
- Zwillinger, D. (1995). *CRC standard mathematical tables and formulae* (30th ed.). Boca Raton, FL: CRC Press.

Index

Entry titles are in **bold**.

- Abbott, Andrew, 769
Abdi, Hervé, 561, 701, 727
Abduction, 1–2
Abell, P., 187
Abraham, B., 398, 399, 402
Abstracts, historical, 463
Abuses, 779
Academy of Management Review, 777
Access, 2–3, 43, 798
 control, 605–606
Accounts, 3–4
ACF (autocorrelation function), 667
Achen, C., 1157
ACT (affect control theory), 306
Action research, 4–6, 19
 participatory, 4–5, 799–800
 team research and, 1114
Action researchers, 606
Active interview, 6–7, 1168
Actor attributes, 719
Actor-Partner Interdependence Model, 291
Actuarial estimator, 453
Adair, John G., 452
Adding Zs method, 637
Additive model, generalized, 418
Adjusted *R*-squared. *See* **Goodness-of-fit measures; multiple correlation; R-squared**
Adjustment, 7–8
ADL (autoregressive distributed lag), 298, 313
Adler, Patricia A., 634
Adler, Peter, 634
Administrative Science Quarterly, 777
Adorno, T., 223–224
AER (average economic regression), 298
Affect control theory (ACT), 306
Afifi, A. A., 1137
African Evaluation Association, The, 340
Agar, Michael, 756
Agent-based simulations, 159
Aggregation, 8, 883, 1157
 See also **Ecological fallacy**
Agresti, Alan, 37, 38, 483, 815, 895
AIC. *See* **Goodness-of-fit measures**
Aiken, Leona S., 500, 691, 695
Ainsworth, Mary, 336
Akaike information criterion
 (AIC), 99, 313, 343, 551, 650
Albritton, Robert B., 518
Aldrich, J. H., 879
Algina, J., 1192, 1193
Algorithm, 9–10, 77, 769
 centroid method, 115–117
Alkin, Marvin, 338–339
Allison, Paul D., 648, 1177, 1178
Allport, Gordon, 567
Alpha, significance level
 of a test, 10–11
Alta Vista, 765
Alternating Least Squares program
 (ALSCAL), 670–671
Alternative hypothesis, 11, 763, 781
Alters, 721
Altheide, David L., 326
Alvesson, Mats, 843, 845
AMELIA, 648
Amemiya, T., 425
American Association for Public Opinion
 Research (AAPOR), 742
American Educational Research
 Association, 178, 179, 181, 182, 213–214
American Evaluation Association, The, 338
American Medical Association, 322
American Psychological Association
 (APA), 322, 337, 645
American Psychologist, 644
American Sociological Association, 332
AMOS, 11, 449, 647
Analog computer simulations, 159
Analysis
 bivariate, 73, 129, 1172
 canonical correlation, 28, 83–86, 794
 categorical data, 29, 96–100, 759
 causal, 92, 151
 cluster, 128–130, 881
 cohort, 140–141
 componential, 157–158
 content, 21, 186–189, 712, 737, 887,
 889–890, 1163, 1165, 1180, 1185

- conversation**, 88, 197–199, 708, 753, 874
correspondence, 71, 203–205, 770
critical discourse, 214–216
discourse, 88, 217, 887, 893
discriminant, 84, 725
factor, 181, 794, 802, 881
 interactional, 707–708
 intervention, 79
latent class, 98, 732
level change, 1130
multilevel, 252, 732, 1173, 1174
multiple regression, 729, 773, 802
multivariate, 881, 1172
narrative, 92, 705–709, 771
network, 719–725
 of Cross-Classified Data Having Ordered Categories, 29
 of means, 13
path, 334, 802–806
 performative, 708
principal components, 71, 84, 205, 725, 855–857
probit, 872–873
qualitative
 meta-, 892
 secondary, 22, 892
 quantitative content, 284
regression, 7, 11–12, 65, 1139, 1179, 1199
 structural, 706–707
 thematic, 186, 706
Tobit, 1134–1135
trend, 1140–1142
units of, 673–674, 1157–1158
univariate, 73, 1137, 1159–1160
verbal protocol, 1180–1181
Analysis of covariance (ANCOVA), 8, 11–12, 74
Analysis of variance (ANOVA), 8, 11, 12–13, 37, 74, 97–98, 124, 321, 390, 541, 593, 625, 638, 652, 762, 910, 1139, 1153, 1172, 1173–1174, 1176, 1195, 1196
 between-group sum of squares and, 65
 fixed- or random-effect models, 16
 history of, 13
 hypothesis testing, 14–15
 magnitudes of effects, 14
 one-way, 12, 13–14, 541, 543, 764–765
 two-way, 13, 15–16, 764, 1149–1151
Analytic induction, 16–18, 91, 1121
 Analytical procedure, 471–472
 Anastasi, A., 182
 Anderberg, M. R., 116
 Anderson, Carolyn J., 348
 Anderson, D. R., 87
 Anderson, N. H., 696
 Andersson, B. E., 218
 Andrews, L., 862
 Angle, John, 489
 Animal behavior. *See* **Ethology**
Annals of Statistics, 77
Anonymity, 18–19, 219
 Anthropologists, cognitive, 305
Anthropology as Cultural Critique, 661
Anthropology of Experience, 661
 APA (American Psychological Association), 322
 Application, fields of, 529–530
Applied qualitative research, 19
Applied research, 19–20
Applied Research Design: A Practical Guide, 20
 Approaches to coding, 289
 Arabie, P., 671
 Arch, D. C., 1180
Archeology of Knowledge, The, 402
Archers, The, 242
Archival research, 20–21, 737
Archiving qualitative data, 21–22
ARIMA (Autoregressive Integrated Moving Average), 22–24, 79–80, 79–81, 259, 313, 397, 439, 667
 Aristotelian logic, 243
 Aristotle, 102–103, 486
 Arithmetic means, 629
 Arithmetic Rule, 410
 Armour-Thomas, E., 227
 Arnold, S., 622
 Arrow, Kenneth J., 24
Arrow's impossibility theorem, 24–25
Artifacts
 in research process, 25–27
 methodological, 452 (*See also* **Artifacts in research process**)
Artificial intelligence, 27–28, 357, 877
 Artificial neural network. *See* **Neural network**
 Ary, D., 752
 Asch, S. E., 451, 483
 Ashby, D., 831
 Ashmore, Malcolm, 404
 Ashworth, C., 767
 AskJeeves, 765
Association, 28–29, 36, 38, 1167
 descriptors, 735
 uniform, 30
 Yule's Q, 1203
Association model, 98
 applications, 32
 estimation, 32–33
 three-way and higher-way table, 31–32
 two-way cross-classified table, 29–31
Assumptions, 33–36, 101, 646, 703, 835, 1154
Asymmetric measures, 28, 36–38

- Asymptotic properties**, 38–39
 Atkinson, P., 137, 661, 708
 Atkinson, Robert, 569
ATLAS.ti, 39, 87
 Atomism, 836
Attenuation, 39–40
Attitude measurement, 40–42
Attitude scales, 711
Attribute, 42
Attrition, 42–43, 157, 782
Auditing, 43
 Australasian Evaluation Society, 338
Authenticity criteria, 43–45, 92, 283
 Authority, 92
 Auto-associators, 727
Autobiography, 45–46
 Autocorrelation. *See* **Correlation**
 Autocorrelation function (ACF), 667
Autoethnography, 46–48, 755
 Autoregression. *See* **Serial correlation (regression)**
 Autoregressive distributed lag (ADL), 298, 313
 Autoregressive time-series, 22, 23
Average, 48–50
 Average economic regression (AER), 298
 Average effect sizes, 637
 Axelrod, Robert, 413, 860
 Axiom, 33
 Ayer, A. J., 586

 Babbage, Charles, 403
 Bacon, Francis, 468, 486
 Baird, Chardie L., 300, 686
 Baker, Frank, 724
 Bakker, Ryan, 432
 Balakrishnan, N., 276
Balinese Character, 1185
 Baltagi, Badi H., 784
 Bandwidth, 422
 Banks, M., 1185, 1186
 Bar chart, clustered, 82 (figure)
Bar graph, 51–52, 1159
 Bargh, J. A., 854
 Barnes, D. B., 389
 Barthes, Roland, 640
 Bartholomew, D. J., 556
 Bartlett, M. S., 86
 Bartunek, J. M., 1115
Baseline, 52
Basic research, 20, 36, 52–53
 Batchelder, W., 139
 Bateson, Gregory, 1185
 Baumrind, Diana, 644
 Bayer, A. E., 1114, 1115
 Bayes, Thomas, 53–54, 55

Bayes factor, 53–54, 57
Bayes Regression, 669
Bayes' theorem, Bayes' rule, 54–55, 859, 1155
Bayesian
 analysis, 433
 estimation, 318–319
 methods, 298
 probability, 412
 statistics, 270
Bayesian inference, 55–58, 612, 650, 857, 1155, 1181
 Bayesian information criterion (BIC), 99, 437, 551
 Bayesian simulation. *See* **Markov chain Monte Carlo methods**
 Baym, N., 1184
 Beck, N., 423, 546, 1134
 Becker, Howard, 18, 150, 461, 799
 Becker, Mark P., 596
Behavior coding, 59
Behavioral sciences, 59
Behaviorism, 60, 887, 1124
 Bell, Susan, 707
Bell Curve, The, 62
Bell-shaped curve, 60–62, 121
 Belmont Principles, 324–325
 Belmont Report, The, 324, 494, 1187
 Benfer, R. A., 358
 Bentham, Jeremy, 1197
 Benton, T., 222
 Benzécri, J. -P., 204
 Benzécri, Jean-Paul, 286
 Berelson, B., 186
 Berlin circle, 585–586
 Bernardo, J., 54, 55
Bernoulli, 62, 723, 801
 Bernoulli, Jakob, 257
 Bernoulli, Johann, 257
 Bernstein, R., 749
 Berntson, G. G., 882
 Berry, W. D., 803
 Bertagnolli, A., 154
Best Linear Unbiased Estimator (BLUE), 62–63, 301, 320, 464, 776, 833, 1154
Beta, 64, 1173
 Bettman, J. R., 1180
Between-group differences, 64, 154
Between-group sum of squares, 14, 65
 Between-method triangulation, 1142
 Beverly, John, 1119
 Bhaskar, Roy, 221, 314, 510, 642
Bias, 65–67, 742, 825, 878, 902, 1120, 1160
 nonresponse, 741–742
 observer, 757–758
 panel data analysis, 783–784

- quota sampling** and, 905–906
 sampling, 825–826
 selection, 901, 1191
See also **Unbiased**
- Biased parameter estimates**, 1160
BIC. *See* **Goodness-of-fit measures**
 Bickman, L., 20
 Biggs, D., 119
 Billig, Michael, 640
Bimodal distribution, 67, 1157
Binary, 67
Binomial distribution, 62, 67–68, 1161, 1207
 negative, 715–716
 Binomial mean and variance functions, 428(figure)
Binomial test, 68–69, 870–871
Biographic Narrative Interpretive Method (BNIM), 69–70
 Biographical method. *See* **Life history method**
Biometrics, 11
Biplots, 70–71
Bipolar scale, 71–72
 Birdwhistle, R. L., 756
Biserial correlation, 72–73
 See also **Correlation**
 Bishop, George F., 774
 Bissett, R. T., 1180
Bivariate analysis, 28, 73, 129, 1172
 Bivariate regression. *See* **Simple correlation (regression)**
 Black, Max, 412
 Black, T. R., 540, 542, 607, 623
 Blaikie, Norman, 1, 243, 310, 378, 456, 469, 487, 510, 767, 786, 837
 Blalock, H., 200, 657, 768, 1147, 1158, 1176
 Blascovich, Jim, 882
 Blau, P. M., 277
Block design, 73–74
 Bloor, M., 633
 BLUE. *See* **Best Linear Unbiased Estimator**
 Blumen, I. M., 666
 Blumer, Herbert, 326
 Blurred genres, 660–661
BMD, 74–75
 See also **Correlation**
BNIM (Biographic Narrative Interpretive Method), 69–70
 Boaz, A., 831
 Bochner, Arthur P., 1198
 Bock, R. Darrell, 770
 Boehmke, Frederick J., 572
 Bogdewic, S. P., 753
 Bohman, James, 839
 Bohrstedt, G. W., 772, 1199
 Boje, D. M., 706
 Bollen, Kenneth A., 449, 739, 802, 901
Bonferroni technique, 10, 75, 263
 Bonnie, R. J., 864
 Bookstein, F. L., 792
 Boolean logic, 410
 “Bootstrap Methods: Another Look at the Jackknife,” 77
Bootstrapping, 39, 75–77, 75–78, 491, 535, 727
 application, 78
 development of, 77–78
 Borgatta, E. F., 1199
 Bornat, Joanna, 48
 Bornstein, Robert F., 368
 Bosworth, M., 21617
 Bottorff, J. L., 753
 Boucke, O. F., 1158
 Boudon’s index, 653
 Boundary-blurring inquiry, 47–48
Bounded rationality, 78–79
 Bourdieu, Pierre, 224, 749
 Bowlby, John, 336
 Box, G. E. P., 22, 55, 79, 80, 81
 Box-and-whisker plot. *See* **Boxplot**
Box-Jenkins modeling, 22, 400, 515, 1128
 ARIMA models, 79–80
 forecasting, 80
 intervention effects, 80–81
 transfer functions, 81
 Box-Steffensmeier, Janet M., 345
Boxplot, 81–82, 278, 543, 1159
 between-group differences and, 65
Boys in White, 799
 Bradburn, N. M., 773
 Bradbury, H., 5, 185
 Brahe, Tycho, 48
 Brainstorming, 445
 Brannick, M. T., 40, 41
 Brazil, K., 896
 Breen, R., 107, 1125
 Breiman, L., 890
 Brennan, Robert L., 419, 420
 Brent, E. E., 358
 Breusch-Pagan tests, 459
 Brewer, M. B., 210, 246, 483, 504
 Bridgman, Percy, 768
 Briggs, Charles, 6
 British government surveys, 235
 British Library National Sound Archive, 21
 Brockmeyer, E., 904
 Bronfenbrenner, Urie, 294
 Bronshtein, I. N., 408
 Brookfield, Stephen D., 377
 Brown, B. K., 641
 Brown, C., 94, 120, 730

- Brown, J., 217
 Brown, Steven R., 887
 Browne, M. W., 208
 Brownie, C., 87
 Brownlee, K., 791
 Brunswick, Egon, 294
 Bryant, C. G. A., 837
 Bryk, A. S., 449
 Bryman, Alan, 242, 252, 254, 264, 269, 325, 337, 378, 402, 461, 483, 485, 577, 633, 655, 681, 698–699, 816, 895, 896
 Buckholdt, D. R., 334
 Buckland, W. R., 321
 BUGS software package, 433
 Bunge, Mario, 104
 Burke, P. J., 589–590
 Burkhart, Ross E., 659
 Burnham, K. P., 87
 Burt, Cyril, 403, 404, 419
 Bury, M., 708

 Cacioppo, J. T., 882
 Cahill, Spencer, 755
 CAIC. *See* **Goodness-of-fit measures**
 Cain, Carole, 706
 Calculating the d-index, 637–638
 Cale, Gary, 377
 Cameron, A. C., 342, 544, 830
 Campbell, Donald T., 154, 229, 337, 354, 356, 502, 503, 515, 516, 618, 640, 678, 712, 819, 899, 1142, 1162–1166
 Canadian Evaluation Society, The, 338
 Canadian Tuberculosis (TB) Project, 5
 Canam, C., 892
 Cancer, analysis of, 632
 Cane, B., 751
 Cannell, C., 59
Canonical correlation analysis, 28, 83–86, 147, 794
 Capability ration, 345
 Capdevila, R., 887
 CAPI. *See* **Computer-assisted data collection**
Capital, 282
 Caplinger, C., 151
Capture-recapture, 86–87, 112, 247
CAQDAS (Computer-Assisted Qualitative Data Analysis Software), 189
 advantages and disadvantages, 88
 program functions, 87–88
 Carless, S. A., 272
 Carlin, B. P., 858
 Carlin, J. B., 433
 Carlson, M., 370
 Carmines, Edward G., 450, 486, 672, 696
 Carrington, P. J., 725
 Carroll, J. D., 671
 Carroll, J. M., 877
Carryover effects, 89, 154, 205
CART (Classification and Regression Trees), 89–90
Cartesian coordinates, 90
 Cartwright, D., 720
 Cartwright, Nancy, 101
Case, 90
 control study, 93–94
 study, 90–93, 420–421, 711, 1144
 Casella, G., 613, 857
 CASI. *See* **Computer-assisted data collection**
 Castellan, N. J., 541, 733, 734, 736
 Casual explanation, 454
 Casual inference, 254
 Catalytic authenticity, 44
Catastrophe theory, 94–96, 730
Categorical, 96
 data analysis, 29, 191, 759
 historical development, 96–97
 missing data in, 99–100
 model comparison and selection, 99
 model fitting and test statistics, 98–99
 types of, 97–98
 predictors, 693
 variables, 609, 895
 Catholicism, Roman, 243
 Cauchy, Augustin, 114
 Causal analysis, 92, 151
Causal mechanisms, 100–101
Causal modeling, 101–102
Causality, 101, 102–105, 439, 837
 probabilistic theory of, 104
 Cavendish, S., 710
 CBR (crude birth rate), 247–248
 CCA. *See* **Canonical correlation analysis**
 CDR (crude death rate), 247–248
Ceiling effect, 106
Cells, 106, 230, 1202
Censored data, 106–107
Censored regression, 1125
Censoring and truncation, 107–108, 108–109, 277, 1161
 analyses of outcomes of, 109
Census, 110–112, 760–761, 766
 adjustment, 112–113
 Center for the Study of Ethics in the Professions, 322
Central limit theorem, 61, 76, 113–115, 281, 1158, 1206
Central tendency, 115, 630, 882, 1159, 1176
Centroid method, 115–117

- CESSDA (European Social Science Data Archives), 234
- Ceteris paribus**, 117
- Ceteris paribus clause*, 295
- CHAID (Chi-Squared Automatic Interaction Detection)**, 118–119
- Chakroaborti, S., 538, 539, 734
- Chamberlayne, P., 48, 69
- Chammah, Albert, 413
- Change scores**, 119–120
- Chaos theory**, 120–121, 730, 850
- Chaotic processes, 258
- Chapin, Tim, 363
- Charlton, John R. H., 245
- Charlton, Tony, 712
- Charmaz, Kathy, 444
- Chartrand, T. L., 854
- Chave, E. J., 40
- Chavez, L., 139
- Chebyshev's inequality, 1113–1114. *See* **Tchebechev's inequality**
- Chell, Elizabeth, 218
- Chen, Peter Y., 40, 41, 415, 502, 513, 527
- Cheng, P. W., 1155
- Chi-square**
- distribution, 631
 - distribution**, 15, 121–123, 542
 - nonparametric statistics** and, 735–736
 - test, 627
 - test**, 75, 123–124, 538, 540
 - Yates' correction**, 1202
- Chiang, A. C., 616
- Chiari, G., 185
- Choice probabilities**, 166
- Chow test**, 124–125
- Christ, Sharon L., 449
- Christians, C., 832
- Ciambrone, Desirée, 327
- Cicerelli, V. G., 615
- Cicourel, Aaron, 6
- Civettini, Andrew J., 259, 667
- Cixious, Hélène, 846
- Clark, James, 707
- Clark, M. H., 615, 822
- Clark, V. A., 1137
- Clarke, Harold D., 80, 439
- Clarke, K. A., 650
- CLASCAL, 671
- Classical inference**, 125–126
- Classification trees. *See* **CART**
- Clatts, Michael, 755
- Clayman, Stephen, 260
- Cleveland, R. W., 1157
- Cleveland, W. S., 422
- Clever Hans (horse), 356–357
- Clifford, J., 845, 1197
- Clinical research**, 127–128
- Clogg, Clifford C., 32, 97, 98
- Closed-ended questions**, 59, 128, 604, 709, 848, 884, 903, 1117, 1168
- Closed settings, 2
- Clough, T. P., 842
- Cluster analysis**, 128–130, 881
- Cluster sampling**, 130, 697
- Clustered bar chart, 82(figure)
- CMLE. *See* **Conditional Maximum Likelihood Estimation (CMLE)**
- Cochrane-Orcutt procedure. *See* **Time-series data**
- Cochran, William G., 11, 96, 130, 821, 901
- Cochran's Q test**, 130–131
- Code. *See* **Coding**
- Code of ethics, 323
- See also* **Ethical codes**
- Codebook. *See* **Coding frame**
- Coding**, 87, 528, 895
- approaches to, 289
 - behavior**, 59
 - contrast**, 193–194
 - determining codes for variables and, 133–135
 - frame**, 136–137
 - levels of measurement, 132–133
 - of responses, 604
 - post**, 135–136
 - pre-**, 848–849
 - qualitative data**, 137–138
 - scheme for **content analysis**, 187
 - theoretical sampling** and, 1121, 1122
 - verbal protocol analysis**, 1180
- Coefficient, regression, 617
- Coefficient of determination**, 138–139, 878
- Coffey, A., 137
- Cognitive anthropology**, 139–140, 158, 305
- Cognitive social structures, 721
- Cohen, A., 512
- Cohen, Jon, 348, 499–500, 637
- Cohen's *d* statistic, 299
- Cohort analysis**, 140–141, 153
- Cointegration**, 141–144
- Cole, Michael, 294
- Coleman, C., 760
- Collier, A., 222
- Collinearity. *See* **Multicollinearity**
- Collins, Randall, 850
- Colson, F., 789
- Column
- association (*See* **Association model**)
 - effects, 30
 - markers, 70

- Commission error, 1180
- Commonality analysis**, 145–147
- Communalities**, 147, 856
- Community advisory committee (CAC), 5
- Community of Those Who Have Nothing in Common, The*, 779
- Comparative method**, 91, 148–151
- Comparative research**, 152–153
- Comparison group**, 153–155
- Comparison operators, 1137
- Competing risks**, 155–156, 344, 1138
- Complete case analysis, 646
- Complex data sets**, 156–157
- Component fit**, 173
- Componential analysis**, 157–158
- Composing Ethnography: Alternative Forms of Qualitative Writing*, 1198
- Computer-assisted data collection**, 159
- Computer-Assisted Telephone Interviewing (CATI)**, 740, 1117–1118
- Computer simulation**, 158–159
- Comte, August, 837
- Concealment, 240
- Concepts**, 101, 716, 1171
- Conceptualization, operationalization, and measurement**, 161–165
- Concurrent validity**, 163
- Conditional independence graph, 439–440
- Conditional likelihood ratio test**, 165
- Conditional logit model**, 166–167
- Conditional Maximum Likelihood Estimation (CMLE)**, 167–168
- Conditional probability, 868
- Conduct of Inquiry, The*, 848
- Confidence intervals**, 76, 97, 168–169, 535, 866, 897
- Confidentiality**, 22, 26, 169, 218, 860–865, 1117
- Confirmability, 43, 1145, 1182
- Confirmatory Factor Analysis (CFA)**, 147, 169–174, 802
- Confounding**, 175–178, 195, 206
- Congressional Record, 765
- Conjugate models, 58
- Consequences, 374, 668
- Consequential validity**, 178–179
- Constant**, 179–180
- comparison**, 180–181, 1121
- Construct**, 161, 181–182, 866
- validity**, 163–164, 182–183, 881, 882, 899, 1171–1172
- Constructing quasi-experiments, 820(table)
- Constructionism**, 91, 749, 887, 1143
- social**, 183–185, 893, 1165
- Constructivism**, 43, 185–186, 214–215, 1144
- Consumption, Keynesian theory of, 296
- Contemporary Cultural Studies, University of Birmingham's Centre for, 307
- Content analysis**, 21, 186–189, 712, 737, 887, 1163, 1165, 1180
- qualitative**, 284, 889–890
- visual research**, 1185
- Content validity**, 163, 1171–1172
- Context effects. *See* **Order effects**
- Contextual questions, 463
- Contextualization, 350
- Contiguity, 345
- Contingency coefficient. *See* **Symmetric measures**
- Contingency table**, 28, 71, 106, 123, 190–192, 205, 435, 772
- Contingent effects**, 192–193
- Continuous distributions, 620
- Continuous variable**, 193, 869, 898
- Contrast coding**, 193–194
- Contrast effect, 774
- Control**, 117, 195–196, 711
- adjustment** and, 7
- group**, 196–197
- Convenience sample**, 197
- Convergent validity**, 164, 1143
- Conversation analysis**, 88, 197–199, 259–260, 332, 708, 753, 874, 1136
- Cook, Thomas D., 154, 229, 515, 899
- Coombs, Clyde H., 287
- Cooper, H., 638–639
- Cooper, J., 484
- Coordinates, 71
- Copp, Martha, 306
- Coppel, D. B., 154
- Copyright, 22
- Corbin, Juliet M., 137, 441, 529, 635, 1123
- Cordova, D. I., 349, 351
- Cormack, R., 86, 87
- Correlation**, 73, 80, 627, 772, 802, 887, 1171
- analysis, canonical**, 28, 83–86, 147, 794
- attenuation**, 39
- biserial**, 72–73
- coefficient**, 174, 202–203, 636, 762
- descriptors, 735
- direction, 200
- dispersion of variables, 201–202
- intraclass**, 1173
- multiple, 28
- multiple correlation coefficient**, 907
- nature of, 199–200
- part**, 788–790
- partial**, 772, 790–792
- serial**, 94
- sign, 200

- statistical potential for generalization, 201
 strength, 200–201
- Correll, S., 1184
- Correspondence analysis**, 71, 203–205, 286, 770
- Cortazzi, M., 706, 710
- Corter, James E., 1140
- Corti, Louise, 236
- Cortina, Jose M., 299
- Costantino, G., 874
- Couch, Carl, 755
- Counterbalancing**, 89, 205–206
- Counterfactual**, 153, 206–207
- Couper, Mick P., 505, 740
- Cousins, B., 800
- Cousins, J. B., 801
- Covariance**, 207–208
 matrix, 172–173
 structure, 208–210, 875
- Covariate**, 97, 210, 1160
- Cover story**, 210
- Covering-law explanation, 358
- Covert research**, 211–212
- Cox, D. R., 342, 453
- Cox, Richard, 1155
- Cox model, 343
- Cox proportional hazard estimates, 345(table)
- Cox test, 650
- Coxon, Anthony P. M., 22, 276, 672
- Crabtree, B. F., 127
- Crafting Selves*, 46
- Cramer, Duncan, 564
- Cramer, J. S., 374, 622
- Cramer's Rule, 256
- Cramér's V. *See* **Symmetric measures**
- Crano, W. D., 210, 246, 253, 483, 484, 546
- Crapanzano, Vincent, 756
- Creative analytical practice**
 ethnography, 212–213, 1197
- Creative interviewing, 1168
- Credibility**, 43, 283, 633, 1136
 trustworthiness criteria and, 1144–1145
- Cressey, Donald, 17
- Cressie-Reed statistic**, 213
- Creswell, J. W., 680
- Criterion-related validity**, 163, 213–214, 1171
- Critical discourse analysis**, 214–216, 810
- Critical ethnography**, 216–217
- Critical hermeneutics**, 217–218
- Critical incident technique**, 218–219
- Critical maximum differences, 540(table)
- Critical pragmatism**, 219–220
- Critical race theory**, 220–221
- Critical ratio**, 173
- Critical realism**, 221–223, 837–838
- Critical theory**, 43, 223–224, 376–377, 893, 1123
- Critical value**, 224–225, 542, 543, 1189, 1194
- CRNI (crude rate of natural increase), 247–248
- Croll, Paul, 751, 752
- Cronbach, L. J., 182, 225
- Cronbach, Lee J., 419
- Cronbach's alpha**, 40, 225–226, 397
- Croon, Marcel A., 609, 611
- Crosby, R. D., 397
- Cross, E. J., 127
- Cross-cultural research**, 226–228
- Cross-lagged**, 228–229
- Cross-sectional data**, 42, 229, 1128
- Cross-sectional design**, 229–230, 834
- Cross-tabulation**, 73, 230, 627, 1199, 1202
- Crossley polling, 905
- Crotty, M., 216, 767
- Crude birth rate (CBR), 247–248
- Crude death rate (CDR), 247
- Crude rate of natural increase (CRNI), 247–248
- Cultural materialism, 305
- Cummings, L. L., 272
- Cumulative frequency polygon**, 230–231
- Curran, P. J., 449
- Current Index to Statistics, The*, 321
- Curtin, R., 742
- Curvilinearity**, 231–232
- CV. *See* **Variation (coefficient of)**
- D-index, calculating the, 637–638
- Dandeker, C., 767
- D'Andrade, R., 158
- Danger in research**, 233–234
- Daniels, H. E., 28
- DAP (Draw-A-Person) test, 874
- Darroch, John N., 439
- Darsee, John, 404
- Darwin, Charles R., 335
- Darwinism, 641
- Data**, 234, 760
 archives, 234–235, 765–766
 censored, 106–107
 cross-sectional, 42, 229, 1128
 ellipses, 1175–1176
 formats, 236
 gross national product, 237
 likelihood, 619
 management, 236–239
 qualitative, 890–891
 missing, 55, 134, 741, 782, 903
 panel, 611, 1131, 1132
 sample, 705, 1151
 sets, 156–157, 699
 theory and, 1123

- time-series**, 22, 23, 1128–1132, 1129, 1172, 1173, 1190, 1194
triangulation, 1142
 truncated, 426
validity, 771
verification, 1181–1182
 Databases, online, 765–766
 David, H. A., 156
 Davidson, James, 313
 Davidson, N., 761
 Davis, James A., 720
 Day, N. E., 554
 de Balzac, Honore, 1197
 de Beauvoir, Simone, 778
 de Boef, Suzanna, 291, 314
 de Condorcet, Marquis, 1197
 de Leeuw, Jan, 286, 548
 de Tocqueville, Alexis, 152
 de Ville, B., 119
 Deacon, D., 681, 699
 Death rates, 248
Debriefing, 239–241
Deception, 240, 241
Decile range, 242
Deconstructionist, 846
Deduction, 243
 Deetz, S., 843
Degrees of freedom, 15, 121, 131, 243–244, 764, 1114
 between-group differences and, 65
 between-group sum of squares and, 64
 Dehejia, R. H., 615, 875
 Dehoaxing, 240–241
 Del Monte, K., 1114
 Delaney, H. D., 718
 Delaney, H. G., 354
 Delanty, G., 314
 DeLeeuw, J., 670
Deletion, 244–245
 Deleuze, Gilles, 846
Delphi technique, 244–245
Demand characteristics, 26, 210, 245–246, 1186
Democratic research and evaluation, 246–250, 340
Demographic methods, 246–247
 DeNavas-Walt, C., 1157
 Denzin, Norman K., 184, 508, 661, 893, 1142, 1143
 Deontological ethics, 323
 Dependability, 43, 1145
Dependability of Behavioral Measurements, the, 419
 Dependence. *See Independence*
Dependent interviewing, 250–251
Dependent observations, 251–252
Dependent variable, 101, 150–151, 252–253, 711, 1124, 1167, 1176, 1195
 (in experimental research), 252–253
 (in nonexperimental research), 253–254
 Y, 52, 763, 1201, 1202
 Derrida, Jacques, 242, 579, 779, 843, 846
 Description, thick, 328
Descriptive statistics. *See Univariate analysis*
 Desensitizing, 240
Design effects, 254–255
Design of Experiments, The, 355
Designs
 block, 73–74
 cross-sectional, 229–230, 834
 experimental, 714, 1181, 1196
 factorial, 1153
 interrupted time-series, 898–899
 longitudinal, 253–254
 nested, 651, 717–719
 quasi-experimental, 901
 repeated-measures, 74, 89, 205, 1196
 research, 16, 26
 total survey, 1135–1136
 unbalanced, 1153
 within-subject, 154, 196, 1196–1197
Determination, coefficient of. *See R-squared*
Determinism, 256–257
Deterministic, 865, 1155
 model, 257–258
 trend, 1140
Detrending, 258–259
 Deviance function, 430
Deviant case analysis, 259–260, 360
Deviation, standard, 60, 64, 76, 260–261, 743, 895, 1119, 1159, 1160, 1176, 1205
 Dewey, John, 848
 Dewey, Thomas, 825
 DFBETAS, 492
Diachronic, 261
Diary, 21, 261–262, 712
 time measurement, 1127–1128
Dichotomous variables, 7, 38, 262–263, 1199
 binomial distribution and, 68
 binomial test, 69
 Dickey-Fuller statistic, 1158–1159
 Difference of means test, 1199, 1202
Difference of proportions, 263
Difference scores, 264
 Digman, J. M., 181
 Dijk, Teun Van, 265
 Dillman, D. A., 822, 902
 Dilthey, Wilhelm, 415, 454–456, 509, 579, 755, 1183
 Diltheyan traditions, 415
Dimension, 264
 Dion, K. K., 451
Disaggregation, 265

Disappearing World, 389

Discourse analysis, 88, 217, 265–269, 887, 893

Discourse and Social Psychology, 265

Discrete, 268–269, 898

choice, 166

Discriminant analysis, 84, 270, 286, 725

Discriminant function, calculation of, 270–271

Discriminant validity, 272

and measurement, 164

Discursive psychology, 266

Discussion protocol, 392

Disengagement, 561

Disordinal interaction, 774

Disordinality, 272–272

Dispersion, 273–275, 1178

Dissimilarity, 275–276

Distance matching, 615

Distribution, 36, 276–280, 1159

bimodal, 67, 1157

binomial, 62, 67–68, 715–716, 801, 1161, 1207

frequency, 1157

geometric, 1161

log-normal, 745

multimodal, 1157

normal, 15, 33, 67, 543, 733, 743–745, 868, 1114, 1145, 1156, 1161, 1167, 1189, 1206, 1207

Pascal, 801–802

Poisson, 68, 98, 802, 828–829, 1137

posterior, 839–841

prior, 55, 840, 857–859

probability, 1154

sampling, 38, 76, 535, 747, 1206

T, 1114

unimodal, 67, 1156–1157

Wishart, 208

Distribution-free statistics, 280–281

Distributions in Statistics, 276

Disturbance term, 281

Docherty, S., 892

Documents, types of, 127, 281–285

Dodendorf, D. M., 890

Dominance data, 287

Double-blind procedure, 26, 246, 285

Douglas, Jack, 1, 1168

Douglass, B., 460

Dovidio, John, 477

Draper, J. A., 746

Dreher, A., 879

Dryzek, J. S., 887

Dual scaling, 286–287, 770

Dummy variables, 11, 289, 605, 1137

Dunbar, J., 20

Duncan, G. J., 783

Dunham, R. B., 272

Dunn MCP, 689

Dupré, John, 101

Duration Analysis. *See* **Event history analysis**

Duration survival, 342

Durbin-Watson statistic, 290–291, 650, 1130

Durkheim, Emile, 152, 293, 487, 641, 642, 643, 837

Durkheim's theory, 243

Durning, D. W., 887

“Dutch Book” arguments, 1155

Duvall, Sue, 388

Dwyer, P., 154

Dyadic analysis, 291, 723

Dynamic modeling. *See* **Simulation**

Easterby-Smith, Mark, 606

Ebel, Robert L., 419

EBSCO, 766

Eckstein, H., 148

Eco, Umberto, 307

Ecological fallacy, 446

Ecological fallacy, 8, 293, 1158

Ecological validity, 294

Econometrics, 142, 295–298

Economic freedom, 1124–1125

ECT (electroconvulsive therapy), 521

Edelman, B., 727

Edelman, M., 742

Edgeworth, Frances, 629

Educative authenticity, 44

Edwards, A., 450

Edwards, Derek, 260, 268, 507

Effect, 30

carryover, 89, 154, 205

ceiling, 106

contextual, 190

contingent, 192–193

fixed, 784, 1133

floor, 106

interaction, 106, 772, 773

intervention, 80–81

log-multiplicative layer-, 32

main, 604–605, 1150, 1153

order, 89, 205, 773–774

Pygmalion, 885–886

random, 784, 1133

size, 298–299

Effects

coding, 299–300

coefficient, 300–301 (*See also* **Path analysis**)

Efficiency, 63, 1154

Efron, Bradley, 77, 78, 535, 857

Egger, Matthias, 388

Ego-centered networks, 721

Egoism, 243

- Eibl-Eibesfeldt, Irenaeus, 336
- Eigenvalues**, 78, 301–302
- bootstrapping** and, 78
- Eigenvector**. *See* **Eigenvalues**
- Eisenberg, L., 127
- Eisenhardt, Kathleen, 685
- Eisner, Elliot, 47
- Elaboration**, 230, 302, 772
- Elasticity**, 303–304
- Elderton, Palin, 122
- Electroconvulsive therapy (ECT), 521
- Electronic recording devices, 346
- Electroshock, 521
- Elementary and Secondary Education Act, U. S., 337–338
- Eliason, S. R., 620, 622
- Elliptical orbit, 468
- Ellis, Carolyn, 47, 212, 755, 1198
- Ellis, Charles H., 574
- Elster, J., 101
- Embedded methods, 815, 895
- Ember, C. R., 226
- Ember, M., 226
- Emden, C., 892
- Emerson, R. M., 753
- Emerson, R. M., 633
- Emic/etic distinction**, 305
- See also* **Componential analysis**
- Emirbayer, M., 101
- Emotions research**, 305–306
- Empiricism, 309
- Empiricism**, 306–307, 838–839, 847, 1123
- Empiricist repertoire, 506–507
- Employment tests, 636
- Empowerment evaluation, 340
- Empty cell**, 307
- See also* **Sparse table**
- Encyclopedia of Biostatistics*, 449
- Encyclopedia of Statistical Sciences, The*, 449
- Endogenous variable**, 102, 308–309, 739, 849, 1149
- Engle, Robert F., 142, 313
- Entropy index (H), 250
- Enumerative induction, 17, 486
- Epistemology**, 227, 306, 309–310, 710–711, 767
- EQS**, 310, 449
- Equal appearing intervals, method of, 1126
- Equivalent degrees of freedom, 422–423
- Erasmus, 465
- Erfahrungen*, 579
- ERFW (error rate familywise), 688
- ERIC, 766
- Ericsson, K. A., 877
- Erikson, Erik, 567
- Erklären*, 454
- Erlang, A. K., 904
- ERPC (error rate per comparison), 688
- Error**, 310–313
- backpropagation, 726
- census**, 112
- commission, 1180
- correction models**, 313–314
- nonsampling**, 742
- prediction, 851–852
- proportional reduction of**, 876–877
- random (*See* **Random error**)
- rate familywise (ERFW), 688
- rate per comparison (ERPC), 688
- serially correlated, 290
- standard**, 97, 535, 797, 897, 1206
- systematic**, 1158
- transcription**, 1136
- transmission, 1180
- type I**, 75, 781, 1151, 1152
- type II**, 1151–1152, 1202
- European Evaluation Society, The, 340
- Escofier, B., 794
- ESM (experience sampling method), 346
- Essed, Ph., 220
- Essentialism**, 314
- Establishing credibility, 633
- Estimation**, 315–319, 1158, 1174
- generalized least squares (GLS), 459
- maximum likelihood**, 9, 30, 56, 78, 80, 166, 167–168, 171, 320, 324, 426, 436, 453, 611, 618–622, 647, 732, 740, 827, 871, 872, 1138, 1174
- of polynomial regression, 833
- parameter**, 243–244, 716, 787–788
- time-series data**, 1129–1130
- transition rate**, 1138
- Estimator**, 38, 319–321
- bias**, 65
- interval, 319
- least squares, 63
- ordinary least squares**, 776–777
- unbiased**, 1154
- Eta**, 321
- Ethical codes**, 4, 321–322, 322–325, 323, 569
- confidentiality**, 169
- covert research**, 211–212
- observational research** and, 756
- unobtrusive methods** and, 1165–1166
- Ethnographers, 633
- Ethnographic content analysis** (ECA), 325–326
- Ethnographic realism**, 326
- Ethnographic tales**, 327

- Ethnography**, 19, 46, 325, 328–332, 709, 737, 752, 771, 893, 1197
creative analytical practice, 212–213
critical, 216–217
 in film, 1185
 meta-, 892
naturalism in, 713–714
organizational, 777–778
postmodern, 841–842
time measurement in, 1128
virtual, 842, 1184–1185
- Ethnomethodology**, 260, 332–333, 749, 756, 875
- Ethnostatistics**, 334
- Ethogenics**, 334–335
- Ethology**, 335–336
- Euclidean geometry, 243
- Euclidean metric, 287
- Euler, Leonhard, 257
- European Social Science Data Archives (CESSDA), 234
- Evaluation**
action research, 5
apprehension, 337
autobiography, 46
 of fit, 171–174
participatory, 800–801
qualitative, 891–892
research, 337–340
utilization-focused, 1169
- Evans, J. L., 227
- Evans, M., 276
- Event-contingent responding, 346
- Event count models**, 341–342
- Event history**, 1160
analysis, 155, 342–346
- Event sampling**, 346–347
- Everitt, B., 128
- Exogeneity conditional likelihood ratio test, 165
- Exogenous variable**, 102, 347, 739, 849, 1149
See also **Endogenous variable**
- Expansion factor, 1191–1192
- Expectancy effect. *See* **Experimenter expectancy effect**
- Expected frequencies**, 98
- Expected value**, 348–349, 828
bias, 65
- Experience, lived, 381
- Experience sampling method (ESM), 346
- Experiment**, 79, 90, 101, 349–354, 875
field, 712, 1164
laboratory, 712
natural, 711–712, 737
 online, 766
- Experimental and Quasi-Experimental Designs for Research*, 618
- Experimental design**, 354–356, 618, 714, 1181, 1196
See also **Experiment**
- Experimental groups, 907–908
- Experimental mortality, 503
- Experimenter expectancy effect**, 26, 356–357
See also **Pygmalion effect**
- Experimenting society, the, 337
- Expert systems** (ES), 357–358
- Explained variance. *See* **R-squared**
- Explanation**, 127, 358, 454, 836
Explanation and Understanding, 148
- Explanatory variable**, 359
- Exploratory data analysis** (EDA), 359–360
- Exploratory Delphi, 245
- Exploratory factor analysis**. *See* **Factor Analysis**
- Extensions, 449
- External validity**, 25, 43, 164, 360–362, 712, 1144
- Extradynamic local structure, 724
- Extrapolation**, 362–363, 363(figure), 850
- Extreme case**, 81, 363
- F distribution**, 15, 365–366
- F ratio**, 366–367
- F*-test, 439
- Face validity**, 331, 367–368
 and measurement, 163
- Factor analysis**, 181, 368–373, 794, 881
 confirmatory, 169–174
 for assessing construct validity, 164
- Factorial ANOVA, 686
- Factorial design**, 374, 1153
- Factorial survey method (Rossi's method)**, 374–376
- Fagerhaugh, S., 1123
- Fail-safe *N*, 388
- Fairclough, N., 215
- Fairness, 44
- Fallacy of composition**, 376
- Fallacy of objectivism**, 376
- Fallibilism**, 376–377
- False feedback, 240
- Falsificationism**, 243, 377–378, 468, 767, 786, 837
- FANI (Free Association Narrative Interview), 404
- Faust, K., 720, 725
- FBI data, 597
- FBI Uniform Crime Reports, 447
- Feasible generalized least squares (FGLS), 425

- Featherstone, M., 843
 Fechner, G. T., 408
Federal Register, 765–766
 Federal Regulations for Protection of Human Subjects of Research, 324
 Feedback, 250
Feminist epistemology, 378
Feminist ethnography, 378
Feminist research, 45, 378–381, 771–772, 1169
 Fenton, N., 681, 699
 Ferrell, J., 216
 Fertility rates, 248
 Fetterman, David M., 332
 Feyerabend, Paul, 839
 Feynman, R. P., 36
 FGLS (feasible generalized least squares), 425
Fiction and research, 382
Field experiments, 383–384, 712, 1164
Field relations, 384–385
Field research, 385–386, 777
 Fielding, N. G., 87, 189
Fieldnotes, 21, 22, 46, 87, 386–386, 753, 1145
 Fields of application, 529–530
 Fienberg, Stephen, 97, 122
File drawer problem, 388
 Fillmore, C., 158
Film and video in research, 389, 1185
 Filmer, Paul, 382
FIML. *See* **Full-Information Maximum Likelihood Estimation (FIML)**
Finite element simulations, 159
 Fink, Bernhard, 336
 Finlay, Barbara, 483
 First-order. *See* **Order**
 Fisher, Ronald, 10, 11, 78, 96, 175, 177, 178, 355, 365, 374, 419, 557, 619, 622, 686–687, 695, 764, 816, 896, 1148
 Fisher-Hayter MCP, 687, 689
 Fisher scoring, 429
 Fisher test, 627
 Fisher's appropriate scoring, 286
Fishing expedition, 389–390
 Fisk, Marjorie, 392
 Fiske, D. W., 640, 678
 Fit, evaluation of, 171–174
Fixed-effects (FE), 784, 1133
 design, 351
 model, 390, 524
 Flaherty, Robert, 389
 Flanagan, J. C., 218
 Fleiss, J. L., 513
Floor effect, 106, 390–391
 Focal persons, 721
 Focus Delphi, 245
Focus groups, 19, 391–395, 604, 712, 887
 method, 395
Focus Groups: A Practical Guide for Applied Research, 20
 Focused comparisons. *See* **Structured, focused comparison**
Focused Interview, The, 392
Focused interviews, 395–396
 Foley, Hugh J., 366
 Fontana, Andrea, 768
Forced-choice testing, 6, 396–397
 Forecasting, 80, 313
 Foreign currency exchange, 406
 Forging, 403
 Formative evaluation, 339
 Formulas, homogeneity analysis, 638
Fortune, 685
 Forward telescoping, 251
 Foster, P., 17, 91, 92, 1124
Foucauldian discourse analysis, 402–403
 Foucault, Michel, 224, 402–403, 778, 842, 846, 847
Foundations of Science, The, 397
 Fowler, Floyd J., 519, 521, 768, 903
 Fowler, J., 59
 Fowler, R. L., 762
 Fox, John, 1175
 Frailty models, 1161–1162
 Frank, I. E., 792
 Frank, Ove, 721, 724
 Frankfort-Nachmias, C., 362
 Franzosi, R., 187
 Fraser, Nancy, 224
 Fraser Institute, 1124
Fraud in research, 403–404
 Frederick, B. N., 145
 Frederick, R. I., 397
Free association interviewing, 69, 404–405, 879
 Free Association Narrative Interview (FANI), 404
 Free-sorting experiment, 669–670
 Freedman, David A., 293
 Freedomia University, 622
 Freeman, D., 639
 Freeman, Linton, 720
 Frequencies, observed, 435
Frequency
 distribution, 405, 1157
 observed, 435, 756–757
 polygon, 405
 relative, 757
 Fretz, R. I., 753
 Freud, Sigmund, 223–224
 Frey, J. F., 768
 Frey, James H., 445, 768

- Friedman, J., 792, 857, 890
 Friedman, Milton, 736
Friedman test, 405–406
 Friedrichs, R., 786
 Friendly, Michael, 665
 Frisch, Karl von, 336
 Fromme, K., 154
Full-Information Maximum Likelihood Estimation (FIML), 167
 Fuller, Duncan, 435
Fuller, W. A., 878
Function, 33, 407–408
 Furbee, L., 358
 Fuss, D., 314
 Future-blind, 70
Fuzzy set theory, 408–412
- Gabriel, K. R., 70
 Gadamer, H-G., 456, 579
 Galilei, 103
 Gallagher, Patricia M., 604
 Gallmeier, Charles P., 562
 Gallup, George, 816, 896, 905
 Galton, F., 589, 624, 629, 695
 Galton, M., 48, 751
 Galunic, D. C., 685
Game theory, 27, 413–414
 Games-Howell MCP, 689–90
Gamma, 414–415, 415, 772
 Gandhi, Mahatma, 567
 Gardner, D. G., 272
 Garfinkel, Harold, 197, 332
 Gasko, M., 1143
 Gatekeepers, 2, 3, 385–386
 Gauss, Carl Friedrich, 61, 745
 Gauss-Markov conditions, 428–429, 560, 574–575
Gauss-Markov theorem, 63, 101, 417, 776
 Gaussian distribution, 743
 GDP (gross domestic product), 295, 312
 Gee, James, 707
 Geertz, Clifford, 328, 662, 841, 1124
Geisteswissenschaften, 415
 Geladi, P., 792
 Gelfand, A. E., 58
 Gelman, A., 433
 Geman, D., 614
 Geman, S., 614
 Gender, 386, 415–416, 632, 814
Gender issues, 416–417
 General laws, 836
General linear model, 102, 417–418, 727
 General Printing Office (GPO), 765
 General Social Survey, 153, 431 (figure), 598, 599, 896
- Generalizability**, 91–92, 216, 374, 418–420, 525–527, 674, 1145
Generalizability Theory, 419
Generalization, 201, 243, 896, 1187
Generalization/generalizability in qualitative research, 420–421
Generalized additive models, 418, 421–423
 Generalized block design, 74
Generalized estimating equations (GEE), 423–424
Generalized least squares, 290, 424–425, 459, 464, 1130, 1133, 1190
Generalized linear models, 178, 341, 426–432, 830, 1190
 Generalized variance inflation factor (GVIF), 1176
 Genres, blurred, 660–661
 Geographic Information Systems (GIS), 238
Geometric distribution, 1161
Geometric distribution, 432
 Geometric mean, 624
 Geometry, Euclidean, 243
 Georges, R. A., 777
 Gephart, Robert P. Jr., 334
 Gergen, K. J., 183
 Gerhardt, Uta, 471, 473
Gerontologist, The, 20
 Gershuny, Jonathan, 1127
 Gibbons, J. D., 277, 281, 406, 681
 Gibbons, Jean D., 538, 539, 734, 1194
Gibbs sampling, 58, 433
 Gibbs sampling algorithm, 613
 Gibson, C. B., 1115
 Gibson, G., 5
 Gibson, Nancy, 5
 Giddens, Anthony, 1, 510, 642
 Gifi, A., 271
 Gigerenzer, G., 1155
 Gilbert, Nigel, 266, 506
 Gill, Jeff, 58, 432, 622, 824
Gini coefficient, 28, 250, 433, 488
 Ginitis, H., 413
 GIS (Geographic Information Systems), 238
 Glaser, Barney, 180, 441, 442, 716, 1121, 1122, 1123
 Glasgow, Garrett, 79, 683
 Glass, Gene, 635, 636
 Glass's index, 653
 Gleser, Goldine C., 419
 GLIM software package, 429
 Gluck, Sherna Berger, 772
 Goal-free evaluation, 339
 Godbey, G., 1127, 1128
 Goethe, J. W., 415

- Goffman, Erving, 197–198, 334, 640, 755
Going native, 799
 Gold, Raymond L., 798
 Goldberg, T. D., 220
 Goldfeld-Quandt tests, 459
 Goldstein, D. G., 1155
 Goldstein, H., 157, 676
 Golub, G. H., 616
 Gomm, R., 17, 91, 92, 1124
 Gompertz model, 453
 Good, B. J., 127
 Good, I. J., 54
 Goodenough, W., 139
 Goodkin, D. E., 154
 Goodman, Janice, 710
 Goodman, Leo, 29, 31, 32, 97, 122, 548, 549, 654, 666
 Goodman-Kruskal's gamma, 415
 Goodman's *RC* model, 30–31
Goodness-of-fit measures, 138, 435–437, 538–539, 873, 878
 Goodwin, C., 198
 Google, 765
 Gopaul-McNicol, S. -A., 227
 Gosset, William, 1114
 Government surveys, 235
 Gower, J. C., 70
Grade of membership models, 437–438
 Graded response model, 532
 Grady, C. L., 792
 Graham, J. M., 147
 Grambsch and Therneau's global test, 343
 Grammer, Karl, 336
 Gramsci, Antonio, 219
 Granada Television, 389
 Granger, Clive W. J., 142, 313
Granger causality, 439
 Granovetter, Mark, 720
 Graph theoretic techniques, 722
Graphical modeling, 102, 439–440
 Graphs, bar, 51–52, 1159
 Gravetter, F. J., 1199
 Gravitare, 468
 Grayson, L., 831
 Green, Donald P., 285, 356, 384
 Green, Paul, 271
 Greenacre, Michael J., 204
 Greenberg, D., 120, 228
 Greene, W. H., 475, 608, 609, 746, 776, 830
 Greenland, S., 178
 Griffin, Howard, 241
 Griffin, L., 151
 Griffiths, W. E., 850
 Griliches, Zvi, 446
 Grinyer, A., 18
 Grizzle, J. E., 481
 Gross, Alan L., 433
 Gross, D., 904
 Gross domestic product (GDP), 295, 312
 Gross national product
 data, 237
 growth model, 650
 Gross reproduction rate (GRR), 249
Grounded theory, 17, 21, 39, 88, 91, 127, 180, 440–444, 606, 706, 747, 748, 771, 799, 812, 892, 1121, 1122, 1123
 ATLAS.ti software for, 39
 NVivo software for, 748
Group interview, 444–445
Grouped data, 4450446
 Grove, J., 1162
 Groves, R. M., 740, 742, 902
Growth curve model, 446–449
 GRR (gross reproduction rate), 249
 Guba, Egon G., 43, 92, 633, 714, 845, 1124, 1144, 1145
 Gubrium, Jaber F., 6, 334, 523, 709, 1168
Guide to Running Surveys and Experiments on the World-Wide Web, A, 766
 Gujarati, Damodar N., 298, 313, 321, 650
 Gunter, B., 712
 Guo, Guang, 319
 Guthrie, A. C., 147
 Guttman, Louis, 42, 286, 450
Guttman scaling, 42, 286, 409, 449–450
 Guttman's coefficient, 547
 Guttman's image analysis, 368
 Haberman, Shelby, 97, 98, 549, 550
 Habermas, Jürgen, 224
 Haddon, A. C., 389
 Hagenaars, Jacques A., 98, 303, 549
 Hagle, Timothy M., 256, 263, 617, 684, 879
 Hald, A., 107
 Halfpenny, P., 837
 Hall, Edward, 756
 Hall, Stuart, 307
 Halliday, Michael, 215
Halo effect, 451
 Halstrom, H. L., 904
 Hamilton, James D., 143
 Hamilton, W. D., 860
 Hamm, M. S., 216
 Hammersley, Martin, 17, 19, 22, 91, 92, 307, 327, 402, 415, 578, 586, 661, 832, 1124
 Han, Shin-Kap, 589
 Hand, D. J., 70

Handbook of Research Synthesis, The, 638

Hanley, W. B., 901

Hannan, A., 712

Hannan, M. T., 156, 453

Harary, F., 720

Harding, S., 750

Hardy, Melissa A., 289, 300, 686

Hare, R. D., 639

Harlow, Harry, 336

Harré, Rom, 314, 335

Harris, C. M., 904

Harris, Marvin, 305

Harvey, A. S., 1127

Harvey, A. C., 401

Harvey, David, 257

Harvey, L., 1168

Hastie, T. J., 423, 857

Hastings, N., 276

Hastings, W. K., 613

Hauser, R. M., 654

Hausman, J., 156, 784

Hawthorne effect, 452

Haxby, J. V., 792

Hay, R. A., Jr., 24

Hayano, David, 47, 711

Hayashi, Chikio, 286

Hayashi's quantification theory, 286

Hayes, S. C., 1180

Hays, W. L., 321, 788, 1147, 1150

Hazard rate, 452–453

Hazards models, 1160–1161

Head Start, 615

Headland, T. N., 305

Healey, J. F., 28

Hebb, Donald, 726

Heckman, James J., 7, 453, 732,

821, 901, 1161

Hedges, Larry, 636, 638

Hedstrom, Peter, 645

Hegel, G. W. F., 415, 641

Heidegger, Martin, 455–456

Heinen, T., 556

Heiser, W. J., 855

Helland, I. S., 792

Hempel, Carl Gustav, 104

Hendrick, T., 20

Hendry, D. F., 296, 313

Henle, M., 211

Henry, Gary T., 825, 906

Henry, N. W., 554, 580

Hepworth, J. T., 901

Herenstein, Richard J., 62

Hermeneutics, 454–457, 893

critical, 217–218

Hess, I., 905

Hesse, Mary, 839

Hessling, Robert M., 391

Heterogeneity, 457

unobserved, 830, 1160–1162

Heteroscedasticity. *See* **Heteroskedasticity**

Heteroskedasticity, 458–459, 715, 1190

Heuristic inquiry, 459–460

Hibbs, D. A., 193

Hierarchical designs, 651

Hierarchical linear models, 674–675,

718–719, 1173

Hierarchical (non)linear models, 252,

460–461, 461, 858

Hierarchy of credibility, 461

Higgins, E. T., 854

Higher-order. *See* **Order**

Hill, Austin Bradford, 821, 901

Hill, R. C., 850

Hilsum, S., 751

Hinchman, L. P., 705

Hinchman, S. K., 705

Hine, C., 1184

Hines, J. E., 87

Hipp, John R., 449

Hippel, Paul T. von, 264, 349

Hirsch, M. W., 730

Hirschfeld, H. O., 286

Histogram, 461–462, 1159

Historical abstracts, 463

Historical methods, 462–464

History, oral, 21, 770–772

Hitler, Adolf, 641

HLM (hierarchical linear model), 449, 674–675

Hoaglin, D. C., 1137

Hobbes, Thomas, 642

Hobert, J. P., 857

Hochberg, Y., 690

Holding constant. *See* **Control**

Holism, 641

Holland, Paul, 720, 821

Hollway, Wendy, 405, 879

Holmes, L. T., 887

Holocaust, Nazi, 644

Holstein, James A., 6, 523, 709, 1168

Holsti, O. R., 186

Holt, D., 157

Homan, Roger, 211, 241, 407

Homans, George, 642

Homogeneity analysis formulas, 638

Homoscedasticity. *See* **Homoskedasticity**

Homoskedasticity, 33, 180, 463–465, 1192

Honneth, Axel, 224

Hopkins, K. D., 718

- Horkheimer, Max, 223
Horn, R., 3
Horrace, William C., 475
Horst, Paul, 286
Horton, N. J., 481
Höskuldsson, A., 792
Hotelling-Lawley trace, 703
Hougaard, P., 457
House, Ernest R., 246
Hout, M., 32
Howell, D. C., 736, 762
Hoyt, Cyril J., 419
Hsiao, C., 783
Hubbell, F. A., 139
Hubble, M. B., 211
Huberman, M. A., 716, 890
Hubert, Larry, 724
Hughes, Everett C., 799
Hughes, J. A., 895
Hughes, K., 1142
Hughes, Rhidian, 1183
Hujala, E., 226
Human nature, understanding, 1182–1183
Humanism and humanistic research, 465–466
Humanistic coefficient, 466–467
Hume, David, 103, 839, 1197
Humean theory, 100
Humphreys, Laud, 211, 755, 799
Hunt, J. C., 879
Hunter, John, 636
Hunter, Ronda, 636
Husserl, E., 579
Hymes, Dell, 707
Hypothesis, 33, 100, 243, 467–468, 1172, 1201
 alternative, 11, 763, 781
 analytic induction and, 16–17
 testing, 104, 1151
 verification, 1181–1182
 See also **Null hypothesis**
Hypothesis testing. *See* **Significance Testing**
Hypothetico-deductive method, 33, 243, 468

I, Rigoberta Menchú, 1119
IBM, 27, 239
Ideal type, 1
Idealism, 473–474, 767
Identification problem, 102, 140, 739
Identity, 778–779
 twenty statements test on, 1147–1148
Idiographic/nomothetic, 892. *See*
 Nomothetic/idiographic
IFDO (International Federation of Data
 Organizations), 234

IIA. *See* Independence of irrelevant alternatives (IIA)
Impact assessment, 475–476
Imperative, The, 779
Implicit Association Test (IAT), 477
Implicit measures, 476–477
Imprinting, 335
Imputation, 477–482, 742
In-depth interviews, 19, 21, 537, 711, 893, 1185
In vivo coding, 528–529
Inclusive evaluation, 340
Income measurement, 49–50, 1133, 1157 (figure)
Identification problem, 474–475
Independence, 482–483
 graph, 439–440
 model, 435
 of irrelevant alternatives (IIA), 166
Independent variable, 101, 252, 347, 483, 541,
 1124, 1135, 1160, 1167, 1173
 in nonexperimental research, 484–485
 Tobit analysis and, 1134
 X, 52, 763, 1199
 Y, 1202
Index, 485–486
 of interrater agreement, 511
Indicator. *See* **Index**
Individual differences scaling (INDSCAL), 671
Individualists, 642
Induction, 243, 486–487
 analytic, 16–18, 91, 1121
Inequality
 measurement, 487–488
 process, 488–490
Inference. *See* **Statistical inference**
Inference, statistical, 38
 using **ordinary least squares (OLS)**
 estimators, 776–777
Inferential statistics, 126, 435–436, 490–491
Inflation, 1191–1192
Influential cases, 491
Influential statistics, 491–493
Informant
 interviewing, 493
 key, 537–538
Informed consent, 43, 493–497, 737, 752,
 753, 766, 1165
Initial coding, 443 (figure)
Inquiry, participative, 606
Insertion sort, 9
Institutional Review Boards (IRBs), 494, 522
Instrumental variable, 497–498
Instrumentation, 503
Insurance liability, 234
Integrated time series, 22, 23
Integration, order of, 143

Intelligence, artificial, 27–28, 357, 857

Intent-to-treat analysis, 177

Intentional fraud, 403

Inter-University Consortium for Political and Social Research, 18

Interaction, 303, 1177

effect, 106, 498–500, 772, 773

model, 605

ordinal, 774–775

types of, 499

variables, 11

See Statistical interaction

Interactional analysis, 707–708

Intercept, 500–501

Intercoder reliability, 187

Interdependence, 291

Intergenerational income mobility, 304

Internal reliability estimates, 501

Internal validity, 25, 43, 164, 355, 502–504, 515, 618, 899, 1144

International Evaluation Association, The, 340

International Evaluation Conference, 340

International Federation of Data Organizations (IFDO), 234

International Network for Social Network Analysis, 720

Internet

search engines, 765, 1165

surveys, 504–505, 766, 902

Interpolation, 505–506

Interpretation, 302, 569

Interpretation of Culture, The, 661

Interpretative repertoire, 506–507

Interpretive biography, 507–508

Interpretive interactionism, 508

Interpretivism, 1, 91, 508–510, 728, 837, 893

Interquartile range, 81, 242, 1159

Interrater agreement, 511–513

Interrater reliability, 513–514, 525

Interrupted time-series designs, 898–899

Interval, 898, 1124, 1159

confidence, 76

estimator, 319

Intervention

analysis, 22, 79, 516–518

basic structures of, 516(figure)

effects, 80–81

planned, 354

Interview, social, 698

Interview guide, 518, 1183

Interview schedule, 518–519

Interviewer-administered questionnaire, 822–823

Interviewer effects, 519–520

Interviewer training, 520–521

Interviewing, 127, 521, 761, 773

active, 6–7, 1168

anonymity in, 18–19

Biographic narrative interpretive method (BNIM) for, 69–70

computer-assisted, 159–161

computer-assisted telephone, 1117–1118

creative, 1168

feminist, 1169

free association, 69, 879

in-depth, 19, 21, 537, 711, 893, 1185

narrative, 709–710

open-ended questions in, 59, 768

oral history, 771

postmodern, 1169

probing and, 871–872

qualitative research, 521–524

semi-structured, 21, 537

structured, 1143

total survey design and, 1135–1136

unstructured, 21, 537, 798, 1168–1169

Intraclass correlation, 513, 524–525, 1173

Intracoder reliability, 525–527

Intrinsic motivation, 350

INUS condition, 104–105

Inverse probability problem, 55

Investigator effects, 527–528

Investigator triangulation, 1142

IQ, 663

Irigaray, Luce, 846

Irrelevant alternatives (IIA), 682

Isomorph, 530

Item response theory, 530–533, 881

Item scores, explained, 529

Iterative weighted least squares (IWLS), 429

Iversen, Gudmund R., 446, 703, 817–818

Iyengar, S., 362

Jaccard, James J., 500, 605, 662, 696, 1150

Jackknife method, 491, 535–536, 727

Jackman, S., 423, 425

Jackson, D. N., 181

Jackson, J. E., 855

Jacobson, L., 885

Jacoby, W. G., 581, 672

James, L. R., 512

James, William, 466, 565

Jameson, F., 843

Jasso, Guillermina, 33, 36, 53, 279, 375

Jefferson, Gail, 198, 875

Jefferson, Tony, 405, 879

Jeffreys, H., 53

Jenkins, G. M., 22, 81

Jenkins, Richard, 386

- Jensen, A., 904
 Jillings, C., 892
 Jin, L., 710
 Johansen, Soren, 142
 John, N. R., 746
 Johnson, D. A., 564
 Johnson, Jeffrey C., 493, 537
 Johnson, Mark, 640
 Johnson, N. L., 276, 277, 279
 Johnson, Paul E., 342
 Johnson, Richard A., 483
 Johnson, Samuel, 1197
 Johnson, T., 767
 Joint plot of respondents and party plans, 288(figure)
 Joliffe, I. T., 855
 Jones, B. S., 345
 Jones, C. R., 854
 Jones, Michael Owen, 777
 Jonhnsen, J. C., 537
 Jöreskog, Karl, 104
Journal of Human Subjectivity, 887
Journal of the American Statistical Association, 29, 112
 Joynson, Robert, 403
 JSTOR, 766
 Judd, C. M., 40
 Judge, G. G., 849
 Judgment-oriented evaluation, 339
Judgments, 322–324
- Kahneman, D., 79, 1155
 Kakkar, P., 1180
 Kalton, G., 59, 783
 Kamin, Leon, 404
 Kan, M., 180
 Kant, Immanuel, 103–104
 Kanto, A. J., 1180
 Kanuha, Valli K., 434, 711
 Kaplan, A., 186, 848
 Kaplan, D., 802
 Kaplan, E. L., 109
 Kaplan, E. L., 342
 Kappa coefficient, 627
 Karuntzos, Georgia T., 585
 Kashy, D. A., 291
 Kasprzyk, D., 783
 Kass, R. E., 857
 Katz, J. N., 1134
 Keat, R., 222
 Kelly, George, 185
 Kendall, David, 904
 Kendall, M. G., 28, 321, 392, 791
- Kendall, P. L., 772
 Kendall's tau coefficient, 414, 627
 Kennedy, Peter E., 827, 1179
 Kenny, D. A., 228, 291
 Kepler, 468
 Keppel, G., 206
 Kerlinger, F. N., 736
 Kernel function, 727
 Keselman, H. J., 690
 Kessler, R., 120, 228
 Key consultant, 537
Key informant, 385–386, 537–538
 Key word counts, 186
 Keynesian theory of consumption, 296
 Khazen, R. S., 901
 Khurana, B., 716
 Kidder, L. H., 40
 Kim, H. S., 882
 Kim, M. P., 1140
 Kinal, T., 878
 Kincaid, H., 713
 Kinder, D. R., 361
 King, G., 93, 341, 622, 828
 Kinsey studies, 262
 Kirk, Roger E., 183, 321, 354, 557, 718
 Kirkman, B. L., 1115
 Kish, Leslie, 230, 538, 1191
Kish grid, 538
 Kivlahan, D. R., 154
 Kleinman, A. L., 127
 Kline, R. B., 802, 803
 Kmenta, J., 475, 849
 Knoke, David H., 589–590, 772
 Knott, M., 556
 Knowledge, mode of, 455
 Koester, Stephen, 755
 Kogan, M., 666
 Kolmogorov distance, 260
Kolmogorov-Smirnov test, 538–541, 627, 734, 736
 continuous population data and, 539–540
 one-sample, 538–539
 Kondo, D. K., 46
 Koocher, G. P., 861
 Korczynski, Marek, 2
 Korean Society for the Scientific Study of Subjectivity, 887
 Kornberg, A., 52–53
 Kotz, S., 276, 277, 279
 Kovtun, Mikhail, 438
 Kowalchuk, R. K., 690
 Kowalski, B., 792
 Kracauer, Siegfried, 889

- Krackhardt, David, 721
 Krauss, Autumn D., 415, 502, 513, 527
 Krieger, S., 842
 Krippendorff, K., 186
 Kristeva, Julia, 779, 846
 Kroonenberg, Pieter M., 71
 Krosnick, J. A., 903
 Krueger, Richard, 395
 Kruskal, J. B., 670
Kruskal-Wallis H test, 541–542, 541–543, 737
 Kuder, G. Frederick, 286
 Kuhn, Manfred, 755, 1147
 Kuhn, Thomas, 749, 785–786, 815, 839, 895
 Kuran, Timur, 850
Kurtosis, 543–544, 746, 773, 787, 1159
 Kuusela, H., 1180
 Kuzel, A. J., 127
 Kvale, S., 521, 879
- Laboratory experiment**, 545–546, 712
 Labov, William, 707
 Lacan, Jacques, 846
 Lachenbruch, P. A., 271
Lag structure, 546–547, 1129
 Lagrange multiplier test. *See* **Time-series data**
 Lahiri, K., 878
 Laird, N., 732
 Lakatos, I., 729, 1182
 Lakoff, George, 640
Lambda, 547
 Lambda, Wilks's, 271
 Lambert, D., 830
 Lanarkshire milk experiment, 355
 Landau, S., 128
 Lane, P. W., 1153
 Langeheine, Rolf, 612
 Langellier, Kristin M., 705, 708
 Langewey, Mary D., 1198
Language, Truth and Logic, 586
 Laplace, Pierre, 54, 61, 114, 258
 Lashley, Karl, 854
Latent budget analysis, 548
Latent class analysis, 98, 549–553, 732
 Latent growth curves, 449
Latent Markov model, 553–554, 612
Latent profile model, 554–555
Latent trait models, 555
Latent variable, 120, 555–556
 in **confirmatory factor analysis**, 169–171, 174
 Latham, G. P., 218
 Lather, Patti, 1198
 Latin American literature, 1118–1119
Latin square, 353–354, 556–557
 Laufer, R. S., 861
 Laurie, Heather, 251, 538
 Lauritzen, Steffen L., 439
 Lavrakas, Paul J., 823
 Law
 of comparative judgment, 1126
 of large numbers, 1113
Law of averages, 557–558
Laws in social science, 558–589
 Lawshe, C. H., 512
 Lax, D. A., 273
 Layder, D., 698
 Lazarsfeld, Merton, 391–392, 549, 554, 580
 Lazarsfeld, Paul, 230, 264, 302, 391–392, 441, 772
 Leahey, Erin, 252
 Leamer, E. E., 296, 298, 650, 878
Least squares, 8, 589–590, 1175, 1200
 estimator, 63
 generalized, 290, 459, 464, 1130, 1133, 1190
 ordinary, 33, 102, 124, 231, 731, 739,
 775–777, 827, 833, 872, 1129, 1154,
 1158, 1179
 regression, partial, 792–795
 two-stage, 1149
 weighted, 1190
Leaving the field, 561–562
 Ledolter, J., 398, 399
 Lee, Alfred McLung, 465
 Lee, H. B., 736
 Lee, Raymond, 87, 189, 234, 1164, 1165
 Lee, Richard, 737
 Lee, T. -C., 850
 Leese, M., 128
 Leiden Group, the, 286
 Leiden University, 271
 Leinhardt, Samuel, 720
 LEM, 666
 Lenski, Gerhard, 488
 Lepper, M. R., 349, 351
 Leptokurtic data, 544
 Leroy, A. M., 275
 Leslie, P. H., 562
Leslie's matrix, 562–563
Level change analysis, 1130
Level of measurement, 163–164
Level of significance, 563
 Level sets, 409
 Levene, Howard, 563, 564
Levene's test, 563–564, 1146
 Levenshtein, Vladimir, 769
 Lévi-Strauss, Claude, 720, 846
 Levin, A., 746
 Levin, I. P., 1150
 Levinas, Emmanuel, 778, 779
 Levine, D. W., 746

- Levine, Robert J., 497, 861
 Lewin, Kurt, 4, 294
 Lewis, David, 207
 Lewis, J., 19
 Lewis-Beck, Michael S., 234, 244, 264, 273, 301, 363, 459, 516, 521, 529, 557–558, 603, 643, 699, 1156, 1167
 Lexis-Nexus, 766
 Li, Zhen, 889
 Liability, insurance, 234
 Liao, Tim Futing, 71–72, 73, 254, 261, 263, 279, 302, 359, 376, 390, 418, 434, 485, 563, 573, 577, 643, 660, 667, 815
Life history interview, 90, 771
Life history method, 564–566
 Life stories, 564
Life story interview, 566–670
Life table, 246–247, 453, 570–572
Likelihood ratio test, 165, 440, 572, 676, 833
Likert scale, 41, 572–573, 627
 Likosky, W., 154
Limdep, 573
 Lin, W. -Y., 1192, 1193
 Lincoln, Yvonna, 43, 93, 184, 633, 661, 845, 893, 1124, 1144, 1145
 Lind, J., 899
 Lindahl, L., 304
 Lindell, M. K., 512
 Lindesmith, Alfred, 17
 Lindley, D. V., 54
 Lindquist, E. F., 419
 Linear congruent generator, 910–911
Linear dependence, 1135
 Linear discriminant analysis, 271
 Linear model generalization components, 429
Linear regression, 38, 143–144, 574–575, 731, 746, 1125, 1155, 1177
 Linear smoothers, 423
Linear structural relations model.
 See **LISREL**
Linear transformation, 575–577, 744
 Linearity estimator, 63
 Lingis, Alphonso, 779
 Lingo, James C., 286
 Linguistics, structural, 382
Link function, 418, 577
 Link tracing studies, 721
 Linnaean taxonomy, 305
 Lipsitz, S. R., 481
LISREL, 174, 449, 577, 647
 Listwise deletion, 244, 646
Literature review, 577–579
Literaturwissenschaft approach, 382
 Little, Daniel, 309, 347, 358, 488
 Little, Jani S., 437, 583
 Little, R. J. A., 478
 Liu, Cong, 673
 Liukkonen, J., 746
Lived experience, 381, 579–580, 711
 Lively, K., 151
 Lix, L. M., 690
 Ljung-Box-Q-statistics, 80
 Loader, C., 423
Local independence, 580–581
Local Knowledge, 661
Local regression, 580–581, 1190
 Locke, John, 838
 Lockridge, Ernest, 1198
 Lockyer, Sharon, 242
 Loess, 422
 Loevinger, J., 182
 Log-likelihood, 619
Log-linear model, 29, 98, 440, 593–596, 723, 759, 772–773
 Log-multiplicative association models, 655
 Log-multiplicative layer-effect model, 32
 Log-normal distribution, 745
Logarithm, 232, 582–584, 833
Logic model, 584–585
 Logical empiricism. *See* **Logical positivism**
Logical positivism, 585–586, 837
Logistic regression, 38, 436, 592, 614, 725, 731, 834
Logit, 426, 834
Logit model, 589–592
 Lohfeld, L. H., 896
 London School of Economics (LSE), 297
 Long, J. Scott, 342, 608, 609, 815, 895
 Longford, N. T., 675
 Longitudinal
 designs, 253–254
 perspective, 261
Longitudinal research, 250, 597–601
 Lord, Frederic M., 286
 Lorencz, B., 127
 Lorenz, E. N., 121
 Lorenz, Konrad, 335, 336
 Lorenz curve, 278–279
 Losch, M. E., 882
 Louis, M. R., 1115
 Louis, T. A., 858
 Love, W. A., 85
 Low-income measurement, 49–50
 LSE (London School of Economics), 297
 LU decomposition, 9
 Lucey, H., 879
 Lundberg, George, 768–769
 Lupia, A., 884

Luther, Martin, 567
 Lütkepohl, H., 850
 Luttrell, Wendy, 710
 Lycos, 765
 Lyman, S., 3
 Lynch, Michael, 333
 Lynn, Peter, 698
 Lyotard, Jean-François, 842, 843, 846

Mplus, 647, 667
 for the SEM approach, 449
 Macaulay, A. C., 5, 127
 MacDonald, R. R., 747
 MacDougall, David, 389
 Mackie, John, 103, 104–105, 151
Macro, 603
 Macroeconomics, 558
 Maddala, G. S., 97, 156
 Madsen, Douglas, 651
 Madwives, 521
 Magidson, Jay, 118, 119, 553, 580
 Magnitudes of effects, 14
 Magnitudes of random errors, 909
 Magnusson, D., 40
Mail questionnaire, 603–604
Main effect, 604–605, 605, 1150, 1153
 Maines, D. R., 39, 216, 1123
 Majority rule, 24
 Malcom, N. L., 151
 Malgady, R. G., 874
 Malinowski, Bronislaw, 434, 711, 1168
Management research, 605–606
 Managerial agendas, 606
Mancova. *See* **Multivariate analysis of covariance**
 Mangione, T. W., 903
Mann-Whitney *U* test, 541, 606–607, 734, 737, 1193, 1195
MANOVA. *See* **Multivariate Analysis of Variance**
 Manski, C. F., 475
 Manstead, A. S. R., 1184
 Manton, K. G., 438, 457, 1161
 Maoris, 507
 MAR (missing at random), 646
 Marcus, G. E., 845, 1197
 Marcuse, Herbert, 219, 223
Marginal effects, 608–609
Marginal homogeneity, 609
Marginal model, 609–611
 Marginal propensity to consume (MPC), 296
 Marginalization, 778–779

Marginals, 611
 Marinez, R. G., 139
 Markers, 70
 Market research, 20
 Markoff, J., 186, 187
 Markov, Andrei A., 611
Markov chain, 611–612
 Markov Chain Monte Carlo (MCMC), 318–319, 433, 479
Markov chain Monte Carlo methods, 58, 101, 612–614, 724
 Marlatt, G. A., 154
 Marquardt, Donald, 1174
 Marquis, K., 59
 Martens, H., 792
 Martin, K., 884
 Martinson, B. C., 1139
 Marx, Karl, 152, 282, 314
 Marxism, 382, 705, 1123
 Mason, Jennifer, 518, 893, 894
Matching, 154, 614–615, 769–770, 826, 875, 906
 Material goods, prototypical distributions of, 278 (figure)
 Materialism, 305, 767
Matrix, 36, 172–174, 615–616, 770
 algebra, 616–617, 1178–1179
Maturation effect, 503, 618
 Matza, David, 326
 Maung, K., 286
 Maxim, Paul S., 491
 Maxim Rules, 410
Maximum likelihood estimation, 9, 56, 78, 80, 312, 316–317, 320, 341, 426, 436, 447, 453, 611, 618–622, 647, 732, 740, 827, 871, 872, 1138, 1174
 association model and, 30
 bootstrapping and, 78
 in **confirmatory factor analysis**, 171
 of conditional logits, 166
 various types, 167–168
 Maxwell, S. E., 354, 718
 Maynard, Douglas, 260, 756
 Maynard, Mary, 381, 417
 MCA. *See* **Multiple classification analysis**
 MCAR (missing completely at random), 646
 McCall, M. A., 901
 McCarthy, P. J., 666
 McCleary, R., 24
 McComber, A., 127
 McCubbins, M. D., 884
 McCullagh, P., 426, 830
 McDaniel, D., 151

- McDonald, J. F., 1134
 McDowall, David, 515
 McElhanon, Kenneth A., 305
 McFadden, D., 156
 McGraw, Kenneth O., 405, 511, 525
 McGrew, W., 336
 McGuigan, Jim, 308
 McIntosh, A. R., 792
 McKelvey, R. D., 879
 McMullin, J. M., 139
 McNemar change test, 627. *See* **McNemar's chi-square test**
McNemar's chi-square test, 622–623, 734, 737
 McPartland, Thomas, 1147
 McPherson, S. O., 1115
 McSweeney, John, 515
 Mead, George Herbert, 1147–1148
 Mead, Margaret, 404, 466, 639
 Meadows, Lynn M., 811, 890
Mean, 48, 623–624, 628–629, 743, 787, 895, 897, 1113, 1114, 1156, 1176, 1199, 1205, 1206
 square, 15
 square error, 624–626, 788
 ***t*-test** of difference of, 1146–1147
 ***t*-test** of single, 1145–1146
 trimming, 1143–1144
 See also **Average; Confidence intervals; Measures of central tendency**
Measure
 mini, 673
 of association, 62, 203, 627
 of central tendency, 60, 628–630, 630, 1199
 ordinal, 775
 repeated, 447
 See **Conceptualization, operationalization, and measurement**
Measurement, 163–165
 complete networks, 720–721
 error, 742
 errors, 163, 174
 invalidity, 163
 levels of, 132–133
 triangulation and, 1142
 validity, 1171–1172
Median, 48, 81, 541, 628, 630–632, 743, 897, 1157, 1159, 1187
 test, 630, 631
 value, 81
Mediating variable, 631–632
 Medicaid, 237
 Mee, A. P., 772
 Meehl, P. E., 182
 Meier, P., 109
 Melody, J., 879
 Melton, G. B., 861
Member validation and check, 562, 633
 See also **Respondent validation**
Membership roles, 633–634
Memos, memoing, 443–444, 634–635, 747, 748
 Menard, Scott, 589, 592, 601, 876
 Mendel, G., 36
 Merkle, D. M., 742
 Merleau-Ponty, M., 579
 Merrett, F., 751
 Merrett, Mike, 251
 Merton, Robert, 391–392, 643, 645, 1123
 Merton Chart of Theoretical Derivation, 35, 36
 Messick, S., 178, 179, 181, 182, 214
Meta-analysis, 635–639, 674, 831
Meta-ethnography, 639, 812, 892
 Metadata, 235
Metaphors, 640, 889
Metaphors We Live By, 640
 Metaphysical behaviorism, 60
 Method of equal appearing intervals, 1126
 Method of maximum likelihood estimation (ML), 297
 Method of paired comparisons, 1126
 Method of reciprocal averages, the, 286
Method of Sociology, The, 17, 466
 Method of successive intervals, 1126
Method variance, 640–641
 Methodological artifact, 452
 See also **Artifacts in research process**
 Methodological behaviorism, 60
Methodological holism, 641–642
Methodological individualism, 642–643
 Methodological triangulation, 1142
Metric variable, 643
 See also **Interval; Levels of measurement**
 Metropolis, N., 613
 Metropolis-Hasting algorithm, 613
 Meulman, J. J., 855
 Meulman, Jacqueline, 286
 Michener, C. D., 115
Micro, 643
Micromotives and Macrobehavior, 101
Microsimulation, 643
 Microsoft, 239
Middle-range theory, 643–644
 Miles, M. B., 716, 890
 Milgram, Stanley, 241, 644
Milgram experiments, 644–645
 Milk experiment, Lanarkshire, 355
 Mill, John Stuart, 91, 92, 100, 104, 175, 415, 468, 642, 643

- Miller, Arthur G., 645
 Miller, D., 327
 Miller, E., 862
 Miller, L., 879
 Miller, W. L., 127
 Millett, Kate, 778
 Milligan, G. W., 116
 Milliken, G. A., 564
Million Random Digits, A, 912
 Mills, C. Wright, 211, 307, 462, 508, 708
 Mincer, J., 408
 Mini measure, 673
 Mische, A., 101
 Mishler, E. G., 706, 707, 709
 Mishra, S. I., 139
 Missing at random (MAR), 646
 Missing completely at random (MCAR), 646
Missing data, 55, 134, 645–646, 741, 782, 903
 in **categorical data analysis**, 99–100
Misspecification, 649–650
 Mitchell, G. E., II, 879
Mixed designs, 650–652
Mixed-effects models, 418
Mixed-method research. *See*
 Multimethod research
Mixture model (finite mixture model), 653
MLE (maximum likelihood estimator), 78, 80,
 312, 316–317, 436, 447
Mobility table, 653–655, 654(table)
Mode, 628, 655, 743
Model, 655–656
Model I ANOVA, 656–657, 910
 Model II ANOVA. *See* **Random-effects model**
 Model III ANOVA. *See* **Mixed-effects model**
 Model misspecification, 290
 Model of mobility, 654(table)
 Model sum of squares, 64
Modeling, 657–659
 threshold effect, 1124
Moderating. *See* **Moderating variable**
Moderating variable, 392, 659–660
 Moderatum generalizations, 420
 Modernist phase, 660
 Modes, 67
 Moeschberg, M. L., 156
 Moffitt, R. A., 1134
 Mohr, D. C., 154
 Molecular score, 673
 Molenaar, I. W., 532
Moment, 660, 773
Moments in qualitative research, 660–662
 Monette, George, 1175
 Monodisciplinary perspectives, 605
Monotonic, 662
 Monotonicity, 116
Monte Carlo simulation, 76, 425,
 662–664, 1134
 Mood, Alexander, 631
 Mooijaart, A., 548
 Mooney, Christopher Z., 78, 664
 Moral sciences, 415
 Moreno, Jacob L., 720
 Morgan, D. L., 679
 Morgenstern, Oskar, 413
 Morley, David, 308
 Morse, Janice M., 127, 890, 1122
Mosaic display, 664–665
 Mosteller, Frederick, 97, 359, 360, 535, 1137
 Mother-infant attachment, 336
 Mount, Timothy, 489
 Moustakas, Clark, 459, 460
Mover-stayer models, 665–666
Moving average, 22, 23, 666–667
 Moynihan, J., 760
 MPC (marginal propensity to consume), 296
 MSE (mean square error), 626
 MTMM (Multitrait-Multimethod), 272
 Muchinsky, P. M., 40
 Mueller, R. O., 802
 Mulaik, S. A., 370, 373
 Mulkay, Michael, 266, 506
Multi-item measures, 673
Multi-method, 712
Multicollinearity, 667–669, 776, 833, 1135, 1179
Multidimensional scaling (MDS), 9, 669–672
Multidimensionality, 672–673
 Multilayer perceptron, 727
Multilevel analysis, 252, 460, 673–677,
 732, 1173, 1174
 Multimethod-multitrait research. *See*
 Multimethod research
Multimethod research, 677–681
Multimodal distribution, 1157
 Multinational Time Use Study, 1127
Multinomial distribution, 681
Multinomial logistic regression model, 156
Multinomial logit, 682–683
 vs. **conditional logit model**, 166–167
Multinomial probit, 683–684
Multiple case study, 684–685
Multiple classification analysis (MCA), 685–686
Multiple comparisons, 75, 686–690
Multiple correlation, 28, 690–691
 coefficient, 907
 See also **Multiple regression analysis**;
 R-squared; **regression analysis**
 Multiple imputation (MI), 647–648
Multiple-indicator measure, 164, 696

- Multiple regression analysis**, 190, 244, 605, 691, 729, 773, 792, 796, 802, 1201, 1202
See also **Regression analysis**
- Multiplicative**, 696–697
- Multistage clustered sample designs, 255
- Multistage sampling**, 130, 697–698, 1173
- Multistrategy research**, 679, 698, 815, 895
- Multitrait-Multimethod (MTMM), 272
- Multivariate**, 699
- Multivariate analysis**, 699–702, 881, 1172
- Multivariate analysis of variance and covariance (MANOVA and MANCOVA)**, 702
- Multivariate matching, 615
- Multivariate modeling, 453
- Multivariate nonlinear regression**, 727
- Multivariate normal distribution**, 744–745
- Murray, Charles, 62
- Musgrave, A., 729
- Myerhoff, Barbara, 706
- N* (*n*), 149, 705
- N6**, 705
- Nabb-Whitney U test, 541
- Nachmias, D., 362
- Naes, T., 792
- Naïve induction, 486
- Namboodiri, K., 615
- Nanda, Harinder, 418
- Nanook of the North*, 389, 1185
- Narayan, K., 711
- Narrative**
 active interviews and, 6
 analysis, 92, 705–708, 771
 encouraging, 709–710
 expression, 69
 interview, 709–710
 otherness and, 778
 testimonio, 1118–1119
- Nash, Chris, 257
- National Assessment of Education Progress (NAEP), 181
- National Commission for the Protection of Human Subjects . . . , 1187
- National Council on Measurement in Education (NCME), 182
- National Crime Victimization Survey, 599
- National Election Studies, 765
- National Long-Term Care Survey, 438
- National Longitudinal Surveys, 686
- National Opinion Research Center, 598–599, 816
- National Opinion Research Center (NORC), 896
- National Research Council for the Social Sciences, 234
- National Youth Survey, 599
- Nationwide, 308
- Native research**, 710–711
- Natural experiment**, 711–712, 737
- Naturalism**, 91, 326, 728, 814
 in **ethnography**, 713–714
 objections to, 713
 scientific, 712–713
- Naturalistic inquiry**, 714–715
- Nazi Holocaust, 644
- Negative binomial distribution**, 341–342, 432, 715–716, 801
- Negative case**, 88, 716–717, 1122
- Negotiation, 3
- Neisser, U., 294
- Nelder, J. A., 423, 426, 429, 830, 1153
- Nelkin, D., 862
- Nelsen, C., 151
- Nelson, F. D., 879
- Nelson-Aalen estimator, 453
- Neo-Kantian traditions, 415
- NESSTAR, 236
- Nested design**, 651, 717–719
- Net reproduction rate (NRR), 249
- Network analysis**, 719–723
- Neufeld, R. R., 20
- Neumann, John von, 413
- Neural network**, 27, 725–726
 architecture, 727
 building blocks of, 726
 learning rules, 726–727
 validation, 727
- Neurath, Otto, 586
- New, C., 222
- Newby, Howard, 387
- Newell, R. W., 750
- Newton, Isaac, 33, 36, 53, 468
- Newton, R. R., 543
- Newton-Raphson algorithms, 551
- Newton-Raphson scoring, 429
- Nichols, J. D., 87
- Nicholson, L., 845
- Nietzsche, Friedrich, 466, 846
- Nilsson, S. G., 218
- Nishisato, Shizuhiko, 286
- Noblit, George, 639
- Nominal-interval tables, 37
- Nominal variable**, 11, 12, 37, 96, 728, 897, 1139, 1159, 1195
 between-group differences and, 64
 between-group sum of squares and, 65
 variation ratio, 1178
- Nominalism, 836
- Nomothetic. *See* **Nomothetic/ideographic**
- Nomothetic/ideographic**, 728–729

- Nonadditive**, 729
- Noninteractional effects, 527
- Nonlinear dynamics**, 94, 120, 729–730
- Nonlinear models, 677
- Nonlinearity**, 731–732
- Nonmonotonicity, 116
- Nonorthogonality, 668
- Nonparametric methods, 631
- Nonparametric random-effects model**, 732–733
- Nonparametric regression. *See* **Local regression**
- Nonparametric statistics**, 414
- chi-square** and, 735–736
- compared to **parametric statistics**, 734 (table), 735
- descriptors of correlation and association, 735
- history of, 736–737
- nature of data and sources for, **733**
- t*-tests** and, 735–736
- tests of significant difference, 733–734
- Nonparticipant observation**, 737–738, 752
- Nonprobability sampling**, 738–739, 905
- Nonrecursive**, 739–740
- Nonrespondents, 740
- Nonresponse bias**, 741–742
- Nonsampling error**, 742
- NORM, 648
- Normal distribution**, 15, 543, 733, 743–745, 868, 1114, 1145, 1156, 1161, 1167, 1206, 1207
- Bayesian theory and, 57
- bell-shaped curve** in, 60–61
- binomial distribution** and, 68
- Kurtosis** and, 543
- multivariate, 744–745
- of error, 33
- standard, 744
- unimodal, 67
- Wald-Wolfowitz test** and, 1189
- Normalization**, 746–747, 1137, 1206
- Norman, D., 720
- Normative Delphi, 245
- Normative statements, 836
- Norpoth, Helmut, 80, 81, 657
- NRR (net reproduction rate), 249
- Nuisance parameters**, 747
- Null hypothesis**, 11, 15, 56, 64, 131, 263–264, 535, 541, 747–748, 763, 781, 797, 1148–1149, 1151–1152, 1189, 1194, 1206, 1207
- Null models, 723–724
- Numeric variable. *See* **Metric variable**
- Nunnally, J. C., 40
- Nuremberg Code, 494
- Nutley, S., 831
- Nuzzo, M. L., 185
- NVivo**, 87–88, 748
- Oakes, D., 453
- Oakley, Ann, 523
- Obedience and authority, studies of, 241
- Obedience experiments, 644
- Objective Knowledge*, 103
- Objective mind, 454
- Objectivism**, 706, 749–750
- Objectivity**, 43, 90, 434, 711, 750–751, 1169
- disciplined, 813–814
- in **observational research**, 754–755
- in **qualitative research**, 893–894
- observer bias** and, 757
- reflexivity** and, 1185–1186
- Oblique rotation. *See* **Rotations**
- Observation**, 127, 711, 898
- non-participant**, 737–738, 752
- outlier**, 779
- participant**, 47, 387, 711, 752, 777, 797–799, 1168, 1184, 1185
- schedule**, 751–752
- structured**, 751, 752, 755
- types of**, 752–753
- Observational research**, 336, 753–756
- Observational samples, weighting for, 1191
- Observed frequencies**, 435, 756–757
- Observer bias**, 757–758
- Odds**, 758–759
- ratio**, 30, 759–760
- Office for Human Research Protection (OHRP), 495
- Official statistics**, 19, 760–762
- Ogive. *See* **Cumulative frequency polygon**
- Ohnuki-Tierney, E., 711
- Oksenberg, L., 59
- Olkin, Ingram, 636
- OLS. *See* **Ordinary least squares**
- Olsen, Randall J., 239
- Olsen, Wendy, 501, 625, 627, 632, 659
- Olsen, Wendy K., 576
- Olshen, R., 890
- Omega squared**, 762–763
- Omission, 1180
- Omitted variable**, 763, 878
- Omnibus fit, 172–174
- Ondercin, Heather L., 301, 362
- One-sided test. *See* **One-tailed test**
- One-tailed test**, 10, 763–764
- One-way ANOVA**, 12, 13–14, 73, 541, 543, 764–765
- Kruskal-Wallis *H* test** and, 541, 543
- Oneal, J. R., 345
- O’Neil, Kevin, 766

- Online research methods**, 765–766, 864, 1165
- Ontological authenticity, 44
- Ontology, ontological**, 767, 786, 837
- Open-ended questions**, 59, 709, 768, 823, 903, 1117
- Open question. *See* **Open-ended questions**
- Open settings, 2
- Open Society and Its Enemies, The*, 314
- Operant Subjectivity*, 887
- Operational definition. *See* **Conceptualization, operationalization, and measurement**
- Operationalization**, 162, 627
- Operationism/operationalism**, 768–769
- Optimal forecasts, 400–401
- Optimal matching**, 769–770
- Optimal scaling**, 770
- Oracle, 239
- Oral history**, 21, 770–772
- Ord, K., 622
- Order**, 772–773
- effects**, 89, 205, 773–774
- Ordinal interaction**, 774–775
- Ordinal measure**, 163–164, 775, 1124
- Ordinal relationships, 627
- Ordinal variables measures**, 37, 1159, 1199
- Ordinary least squares (OLS)**, 33, 67, 102, 124, 231, 312, 315–316, 320, 341, 424, 464, 497, 586, 731, 739, 775–777, 827, 833, 872, 1129–1131, 1154, 1158, 1179
- Organizational ethnography**, 777–778
- Origins, 509
- Orne, Martin T., 26
- Orthogonal rotation. *See* **Rotations**
- Osgood, C. E., 41
- Oshina, T. C., 1192, 1193
- Other, the**, 46, 778–779
- Otis, D. L., 87
- Outhwaite, William R., 222, 314, 510, 837
- Outliers**, 81, 232, 779, 897, 1159
- Overparameterization, 474
- P value**, 10, 15, 76, 764, 781, 897, 1114, 1151, 1189–1190, 1194
- between-group differences** and, 64
- Paccagnella, L., 1184
- PACF (partial autocorrelation function), 667
- Page, Stewart, 1164
- Pagès, J., 794
- Paired comparisons, method of, 1126
- Pairwise deletion**, 782
- Pairwise correlation.
- See* **Correlation**
- Paluck, Elizabeth Levy, 285
- Pampel, Fred C., 360
- Panel**, 782
- data, 611, 1131, 1132
- data analysis**, 782–785
- surveys, 42, 250
- Pankratz, L., 397
- PANMARK, 666
- Papineau, D., 712
- Paradigms**, 698, 785–787, 1169
- argument, 815, 895–896
- Paradis, G., 127
- Parameter**, 55, 315, 618, 725, 731, 743, 763, 787, 835, 1154, 1155
- bias**, 65, 1160
- control, 92
- estimation**, 168, 170–172, 243–244, 716, 787–788
- logistic models, 531–532
- models, variable**, 1173
- nuisance**, 747
- Parametric statistics compared to nonparametric statistics, 734 (table), 735
- Parametric test, 622
- Parcerro, Osiris Jorge, 304
- Park, Robert, 798
- Park, Sung Ho, 498, 506, 575
- Parker, Ian, 242
- Parkin, Frank, 307
- Parkinson, B., 1184
- Parks-Kmenta model, 1134
- Parsons, Talcott, 53, 641, 643
- Part correlation**, 788–790
- Partial autocorrelation function (PACF), 667
- Partial correlation**, 772, 790–792
- Partial least squares regression**, 792–795
- Partial regression coefficient**, 795–797
- Partial unstandardized regression coefficient**, 1166–1167
- Participant**
- artifacts, 26
- observation**, 47, 387, 711, 715, 752, 777, 797–799, 1168, 1184, 1185
- volunteer subjects**, 26
- Participative inquiry, 606
- Participatory action research (PAR)**, 4–5, 799–800
- Participatory evaluation**, 800–801
- Partington, M. S., 901
- Pascal distribution**, 432, 801–802
- Patai, Daphne, 772
- Paterson, B. L., 892
- Path analysis**, 102, 334
- data-model fit assessment, 804–806
- foundational principles, 802–804
- Path diagram**, 170–171, 439

- Patton, Michael Q., 47, 340, 460, 476, 715, 716, 801, 812, 891
- Peacock, B., 276
- Pearl, J., 901
- Pearson, Karl, 96, 122, 123, 618, 636, 695
- Pearson correlation coefficient**, 275, 321, 627, 782, 907
- Pearson-Type distributions, 619
- Pearson's chi-square test, 348
- Peirce, Charles S., 1, 847
- Pelusi, J., 127
- Peng, Chao-Ying Joanne, 73, 74, 354
- Pen's Parade, 276
- Peracchio, L., 712
- Percentile**, 630–631, 1206
- Percival, Thomas, 322
- Percy, W. J. L., 901
- Performative analysis, 708
- Period effects**, 140
- Perks, R., 770, 771
- Personal construct psychology (PCP), 185
- Personal digital assistants (PDAs), 346
- Personal narratives, 45–46
- testimonio**, 1118–1119
- Personal relations, 2
- Perspectives, monodisciplinary, 605
- Peters, M. E., 361
- Peterson, Eric, 708
- Petty, R. E., 882
- Pfungst, Oskar, 25
- Phenomenalism, 836, 837
- Phenomenology**, 88, 460
- Phenylketonuria (PKU), 821
- Phi coefficient**, 627, 772
- Phillips, Nelson, 217, 403
- Phillips, Peter C. B., 142
- Philological hermeneutics. *See* **Hermeneutics**
- Philosophical naturalism, 712
- Physical attractiveness stereotype, 451
- Piaget, Jean, 185, 257
- Pickering, Michael J., 810, 889
- Pie chart**, 51, 1159
- Pierce, J. L., 272
- Pike, Kenneth L., 305
- Pillai-Bartlett trace, 703
- Pilot study**, 823–824, 853
- Pimentel, E. E., 32
- Pink, Sarah, 389, 1185, 1186
- Placebo**, 154, 824
- Planned intervention, 354
- Plato, 641
- Platt, Lucinda, 82, 624, 630, 631
- Playfair, Charles, 52
- Plummer, Ken, 45, 281, 466, 467, 566
- Poincare, Henri, 397
- Poisson distribution**, 68, 98, 341, 802, 828–829, 1137
- Poisson regression**, 426, 829–830, 1134
- Poland, Blake D., 1136
- Policy-oriented research**, 830–831
- Polish Peasant in Europe and America, The*, 466, 467, 564
- Political action committee (PAC), 344
- Politics of research**, 831–832
- Polling, 153, 825, 905
- Pollins, Brian M., 465, 650
- Pollner, M., 633
- Pollock, K. H., 87
- Polygon**, 832
- Polynomial equation**, 832–834
- Polytomous variable**, 834
- Pooled surveys**, 834
- Popkin, S. L., 884
- Popper, Karl, 33, 103, 104, 126, 314, 376, 467, 468, 641, 642, 839, 850, 1182
- Population**, 68, 76, 197, 246, 541, 733, 747, 761, 782, 834–835, 894, 905, 1115, 1135, 1151, 1200, 1205
- Beta** and, 64
- bias**, 65
- constitution of cases and, 148–149
- parameter**, 847
- pyramid**, 247, 835–836
- sample size, 705
- See also* **Census**
- Porter, S., 222
- Positivism**, 91, 217, 486, 710–711, 713, 728, 767, 768–769, 836–837, 895, 1144
- Post hoc comparison. *See* **Multiple comparisons**
- Postcoding, 135–136
- Postempiricism**, 838–839
- Posterior distribution**, 56, 839–841
- Postmodern ethnography**, 841–842
- Postmodernism**, 43, 381, 661, 842–846
- interviewing, 1169
- Postmodernist literature reviews, 578
- Poststructuralism**, 751, 778–779, 846–847, 1121, 1183, 1198
- Potential comfounds, 503
- Potential outcome model, 175–176
- Potter, Jonathon, 265, 507
- Potvin, L., 127
- Poundstone, W., 860
- Power, M., 843
- Power efficiency, 733
- Power of a test**, 847
- Power transformations. *See* **Polynomial equation**
- Powers, Daniel, 815, 895
- Pragmatism**, 376, 847–848

- Prais-Winsten procedure, 425
- Precoding**, 848–849
- Predetermined variable**, 849–850
- Prediction**, 387, 827, 850
 equation, 851–852
- Predictive validity**, 163
- Predictor, regression model, 605
- Predictor variables**, 699, 725, 852–853, 1175, 1177, 1202
- Predictors, 699
- Premise, 33
- Presentation of Self in Everyday Life, The*, 640
- Presser, S., 128, 742, 773, 903
- Pretest**, 741, 853–854
 behavior coding and, 59
 group, 604
 sensitization, 854
- Price, Laurie, 737, 755
- Price, Richard, 54, 55
- Primacy effect, 774
- Priming**, 854–855
- Principal components analysis**, 71, 84, 205, 669, 699–700, 725, 727, 792, 855–857
- Principia*, 36
- Principle coordinates, 71
- Prior distribution**, 55, 840, 857–859
- Prior probability**, 859, 1155
- Prisoner's dilemma**, 413, 859–860
- Privacy and confidentiality**, 26, 860–865
- Probabilistic**, 865–866
 reasoning, 358
- Probabilistic Guttman scaling**, 866
- Probability**, 54, 67, 76, 87, 715, 781, 787, 866–869, 1152
 and random numbers, 911–912
 Bernoulli, 62
 conditional logit model, 166
 density function, 56, 65, 828, 869–870, 1206
 distribution, 1154
 prior, 859, 1155
 sampling, 603, 826, 905, 1191 (*See also* **Random sampling**)
 uncertainty and, 1154–1155
 value, 870–871
- Probing**, 522, 871–872
- Probit analysis**, 426, 872–873
 and CMLE, 168
 and **conditional likelihood ratio test**, 165
- Proc Mixed in SAS, 449
- Process use, 339
- Proctor, C. H., 866
- Projective techniques**, 19, 873–874
- Proof procedure**, 874–875
- Propensity scores**, 875–876, 901
- Proportional reduction of error (PRE)**, 876–877
- Proportionate stratified sampling, 826
- Prosch, Bernhard, 413
- Protestant Ethic and the Spirit of Capitalism, The*, 282
- Protestantism, 243, 293
- Protocol**, 604, 877
- Prototypical distributions of material goods, 278(*figure*)
- Provost, Fred, 269
- Proximity relationships, 1140
- Proxy reporting**, 877–878
- Proxy variable**, 165, 878
- Pryce, Ken, 241
- Pseudo-panels, 785
- Pseudo-R-squared**, 878–879
- Pseudo-random numbers, 910–911
- PsychExperiments, 766
- Psychoanalytic methods, 879–880
- Psychological Record*, 887
- Psychology, discursive, 266
- Psychometrics**, 880–881
 Psychometrika, 287
- Psychophysiological measures**, 881–883
- Public opinion research**, 883–884
- Public settings, 2
- Purposive sampling**, 884–885, 893
- Purvis, Lynne M., 396
- Pygmalion effect**, 885–886
- Pyramid, population, 247
- Q methodology**, 887–888
 Q-Methodology and Theory, 887
- Q sort. *See* **Q-methodology**
- Q-sort technique, 396
- Quadratic assignment procedure (QAP), 724
- Quadratic equation**, 888–889
- Qualidata Centre, 22
- Qualitative content analysis**, 889–890
- Qualitative data**
 archiving, 21–22
 coding, 137–138
 management, 810–811, 890–891
 research, 21
 secondary analysis of, 22
- Qualitative evaluation**, 811–812, 891–892
- Qualitative meta-analysis**, 812, 892
- Qualitative research**, 17, 46, 379, 440, 633, 660, 678–679, 711, 748, 813–814, 892, 893–896, 1122, 1124, 1136, 1142, 1183, 1186
- Qualitative semiotic approaches, 284
 See also **Qualitative content analysis**

Qualitative variable, 814, 894–895

Qualrus, 87

Quantile, 815, 895, 1159

Quantitative and qualitative research,

debate about, 815–816, 895–896

Quantitative content analysis, 284

Quantitative research, 711, 742, 816–817, 895–897, 1122, 1142

Quantitative variable, 11, 12, 64, 65, 817–818, 894, 897–898, 907, 1139, 1195

Quartile, 818, 898

Quasi-experiment, 73, 154, 196, 818–822, 898–902

Quenoille, M., 535

Questionnaires, 711, 761, 822–823, 887, 902–904, 1162

self-administered, 773, 895, 902

Questions

closed-ended, 59, 128, 604, 709, 848, 884, 903, 1117, 1168

open-ended, 59, 709, 768, 823, 903, 1117

Quetelet, Adolphe, 48, 61, 629

Queueing theory, 824, 904–905

Quinn, Kevin M., 614

Quota sample, 738, 771, 825–826, 905–906

R (Multiple correlation coefficient), 907

R (Pearson correlation coefficient), 907

R-squared, 435, 650, 762, 1177

Rachlin, Howard, 60

Radbill, L., 251

Radcliffe-Brown, A. R., 720

Radial basis function (RBF), 727

Raghunathan, T. E., 481

Ragin, C. C., 149, 150, 151

Raising, 1191–1192

Rajaratnam, Nageswari, 419

Ramsey, F. P., 1155

Rand, Ayn, 523

RAND Corporation, 244–245

Random assignment, 355, 361, 535, 618, 766, 816, 896, 907–908, 1149

Random-digit dialing (RDD), 505, 603, 825, 884, 905, 908, 1117

Random effects (RE), 784, 1133

model, nonparametric, 732–733

Random error, 908–909

Random factor, 910

Random number generator, 910–911

Random number table, 911–912

Random numbers, 1137

Random sample, 65–66, 76, 93, 738, 747, 835, 896, 897, 1151, 1154, 1194

Random sampling, 626, 817, 912–913

Random variable, 39, 65, 179, 296, 715, 787, 1113, 1155

Random walk, 80

Randomized-blocks design, 73–74, 74, 651, 916–917

Randomized control trial (RCT), 127, 831

Rank order, 898

Rao, C. R., 367

Rapoport, Anatol, 413, 720

Rasch scaling, 409

Rasinski, K., 823, 903

Rassler, Susanne, 481

Rate, transition, 1138–1139

Rathborn, J. C., 901

Ratio

odds, 30, 759–760, 1159

variation, 1178

Rationalists, 309, 847

Rationality, bounded, 78–79

Raudenbush, S. W., 449

Reactivity, 3, 26, 211, 1164

Read, C. B., 277

Reading Auschwitz, 1198

Realism, 326, 713, 749, 767, 1143

critical, 221–223, 837–838

Realist evaluation, 340

Reason, P., 5, 185

Recency effect, 774

Reciprocal averages, method of, 286

Recording devices, electronic, 346

Records of analysis, 634

Recursive, 102

Red bus/blue bus paradox, 166

Redfield, Robert, 281

Redundancy, 392

Reed-Danahay, D. E., 711

Reference axes, 372

Referential content analysis, 186

Reflexivity, 45, 380, 813, 841, 893, 1184, 1185–1186

Reformation, The, 243

Regression, 52, 63, 73, 763, 774, 818, 1128, 1158, 1172, 1194

analysis, 7, 11–12, 65, 101, 258, 367, 608, 729, 1139, 1179, 1199

censored, 1125

coefficient, 617, 779, 795–797, 905, 1114, 1125, 1151, 1174

linear, 38, 143–144, 731, 746, 1155, 1177

local, 1190

logistic, 38, 725, 731, 834

multiple, 190, 792, 802, 1201, 1202

multivariate nonlinear, 727

negative binomial distribution, 715

partial least squares, 792–795

- Poisson**, 829–830, 1134
 polynomial, 833
quantile, 895
 robust, 1190
seemingly unrelated, 739
 simple, 1201, 1202
 Reich, J. W., 901
 Reich, Willhelm, 224
 Relational database software, 239
 Relational ties, 719, 721
Relationships, 1201
 nonlinearity of, 731
 tree diagrams for representing, 1140
 Relative distribution method, 279
Relative frequency, 757, 859
Relativism, 749, 751, 1183
 text meaning and, 1120
Reliability, 43, 164, 216, 711
 analysis, 342
 attenuation correction for, 39–40
 in **observational research**, 755
 intercoder, 187
 key informant, 537
 measures, 525
 test-retest, 1119–1120
 true score and, 1144
 Renninger, LeeAnn, 336
Repeated-measures design, 73–74, 74, 89, 205, 447, 1196
 Repertoire, 506–507
 Replication, 26
Representative samples, 712, 733, 761, 883
 Representativeness, 283, 603
 Rescher, N., 751
Research
 action, 4–6, 19, 1114
 participatory, 4–5, 799–800
 applied, 19–20
 applied qualitative, 19
 archival, 20–21, 737
 artifacts, 25–27
 basic, 20, 36, 52–53
 clinical, 127–128
 comparative, 152–153
 covert, 211–212
 cross-cultural, 226–228
 design, 16, 26
 ethnographic, 19
 field, 777
 longitudinal, 250
 market, 20
 methods of, 379
 multistrategy, 679, 698, 815, 895
 native, 710–711
 observational, 336, 753–756
 policy-oriented, 830–831
 politics of, 831–832
 public opinion, 883–884
 qualitative, 17, 21, 711, 892, 1122, 1142, 1183, 1186
 quantitative, 711, 742, 896–897, 1122, 1142
 survey, 1115
 team, 1114–1115
 visual, 1185–1186
 women as subjects of, 45, 771–772, 1169
 RESET test, 650
Residual, 8, 172–173, 174, 1194
Residual sum of squares, 65, 124–125, 1196–1197
 See also **Sum of squared error**
Respondent validation, 633
Respondents, 905, 1135
 non, 740–741
 Response coding, 604
 Response-order effect, 773–774
Response rates, 1115
 Responsive evaluation, 340
 Retrieval bias, 388
 Rex, J., 510
 Rex, John, 1
 Rholes, W. S., 854
 Richards, L., 890
 Richardson, Laurel, 47, 212, 522, 1198
 Richardson, Marion, 286
 Ricoeur, Paul, 217, 778
Ridge-Regression, 669
 Riessman, Catherine Kohler, 708, 709
 Rigdon, E. E., 739
 Riggs, P. J., 786
 Riles, Gilbert, 328
 Rips, L. J., 823, 903
 Risher, Ronald A., 286
Risks, competing, 155–156, 344, 1138
 Ritchie, J., 19
 Ritzer, George, 645
 Robert, C. P., 613
 Robins, J. M., 177, 178
 Robinson, J. P., 1127, 1128
 Robinson, W. P., 17
 Robinson, W. S., 150, 293, 446
 Robson, C., 752
 Robust regression, 1190
 Rochford, E. B., 561
 Rodgers, J. L., 535, 536
 Rodríguez, G., 457, 1161
 Rog, D. J., 20
 Roman Catholicism, 243
 Romney, A. K., 139

- Ronana, W. W., 218
 Roosevelt, Franklin D., 825
 Root mean squared error (RMSE), 852
 Roper studies, 816, 896, 905
 Rose, G., 1185
 Rosen, B., 1115
 Rosenau, P. M., 845
 Rosenbaum, Paul, 821, 875
 Rosenberg, Milton J., 26
 Rosenberg, Morris, 230
 Rosenberg, S., 1140
 Rosenberger, J. L., 1143
 Rosenbluth, A. W., 613
 Rosenbluth, M. N., 613
 Rosenthal, G., 69, 154
 Rosenthal, Robert, 25, 26, 245, 357, 388, 528, 564, 636, 637, 861, 885, 1162, 1186, 1187
 Rosenzweig, Saul, 25
 Rosnow, Ralph L., 25, 154, 564, 1186, 1187
 Rossi, Peter H., 375
 Rothamsted Experimental Station, London, 557
 Rothman, K. J., 93, 94, 175, 177, 178
 Rothstein, Hannah R., 388
 Rousseau, P. J., 275
 Row effects, 30
 Row markers, 70
 Roy's greatest characteristic root, 703
 Rubenstein, A., 78
 Rubik's cube, 353
 Rubin, D. B., 433, 479, 480, 481, 636, 821, 875, 876
 Rudas, Tamás, 558
 Rudestam, K. E., 543
Rules of Sociological Method, The, 487
 Ruskey, F., 1179
 Russell, Bertrand, 102
 Russett, B., 83, 85, 345
 Rustin, M., 69, 879
 Ryle, Gilbert, 1124
- Saad-Haddad, C., 127
 Sacks, Harvey, 197–198, 332, 875
 Safety zone, interactional, 233
 Sale, J. E. M., 896
 Salem, A. B. Z., 489
 Salmon, Wesley, 100–101
 Samoa, ethnography of, 639
Sample, 62, 67, 894, 908–909, 1114, 1127
 binomial test, 68
 convenience, 197
 data, 1151
 observational, 1191
 quota, 738, 771, 905–906
 random, 65–66, 76, 93, 747, 835, 896, 1151, 1154, 1194
 representative, 712, 733, 761, 883
 size, 705
 stratified, 1192
Sampling, 866
 cluster, 130
 content analysis, 187
 distribution, 606–607
 distribution, 38, 76, 168–169, 535, 747, 1206
 error, 742
 fraction, 906
 Gibbs, 58, 433
 Kish grid in, 538
 methods related to **quota sampling**, 906
 multistage, 130, 1173
 natural experiment, 712
 nonprobability, 738–739, 905
 probability, 1191
 purposive, 884–885, 893
 random-digit dialing, 1118
 simple random, 897
 snowball, 771, 884
 theoretical, 884, 1121–1122
 Sandelowski, M., 812, 814, 892
 Santos, F., 685
 Sapir, E., 894
 Sapsford, Roger, 462
 SAS, 236, 239, 621, 648
 SAS macros, 603
Saturation, theoretical, 1122–1123
 Saunders, P. T., 94
 Sayer, A., 222, 314
 Scalar measures, 436
Scale, 673
 bipolar, 71–72
 scores, 903
Scaling
 optimal, 770
 probabilistic Guttman, 866
 Thurstone, 40–41, 1125–1127
Scatterplot, 73, 203(figure), 278–279, 316(figure), 360, 703, 1139, 1199, 1201, 1202
 Schafer, J. L., 433, 479
 Schafer, W. D., 75
Schedule, observation, 751–752
 Scheffé, Henry, 687
Scheffé's test, 765
 Schegloff, E. A., 197–198, 875
 Schegloff, Emanuel, 260
 Schelling, Thomas, 101
 Schenker, Nathaniel, 481
 Schleiermacher, Friedrich, 454
 Schmidt, Frank, 636
 Schmidt, P., 475

- Schmidt, Tara J., 391
 Schneider, H., 107
 Schölkopf, B., 727
 Schrodt, Philip A., 258
 Schuman, H., 128, 773, 903
 Schutz, Alfred, 509, 510
 Schütz, Alfred, 1, 456, 466
 Schwartz, R., 1142, 1162–1166
 Schwarz, N., 773
 Schwarz information criterion (SIC), 313
 Science Aptitude Test, 72
 Scientific naturalism, 712–713
 Scope dimension, 476
Scores
 change, 119–120
 scale, 903
 true, 1144
 Scott, John, 285, 725
 Scott, M. B., 3
 Scriven, Michael, 105, 338
 Seam effect, the, 251
 Sechrest, L., 1142, 1162–1166
Second Sex, The, 778
Secondary analysis of qualitative data, 22, 892
Seemingly unrelated regression, 425, 739
 Segrè, E., 53
 Selden, S. C., 887
Selection bias, 453, 901, 1191
 Selection-history interactions, 503–504
Self-administered questionnaire, 42, 773, 895, 902
 Self-discovery, 778–779
Semantic differential scale, 41
 Semendyaev, K. A., 408
Semi-structured interviews, 21, 537
Semiotics, 217, 382, 810
Sensitive topics researching, 1183
Serial correlation, 94, 290
Sex and Temperament in Three Primitive Societies, 466
 Shadish, William R., 154, 229, 615, 822, 899, 901
 Shaffir, William, 385
 Shapiro, G., 186, 187
 Shaw, L. L., 753
 Sheather, S. J., 1143
 Shepard, Roger N., 670
 Shipan, Charles R., 656
 Shively, W. P., 1157
 Shrout, P. E., 513
 Shuttleworth, J., 879
 Si, M., 20
 Sieber, Joan E., 241, 322, 325, 497
 Siegel, J. S., 250
 Siegel, S., 541, 543, 733, 734, 736
 Signal-contingent responding, 346
Significance
 level, 764, 1148
 of coefficients, 38
 tests, 56, 75, 85–86, 168, 299, 789, 792, 847, 1202, 1206
 alpha, 10–11
 Sijtsma, Klaas, 533, 555
 Sikkel, D., 548
 Silverman, D., 708
 Simmel, Georg, 755
 Simon, B., 751
 Simon, Herbert, 78, 774, 877
 Simple regression model, 1201, 1202
 Simplex models, 369
 Simpson's Paradox, 293, 446
 Simulated annealing, 9
Simulation, 76, 158–159
Simultaneous equation, 102, 475, 739, 746, 849
 Simultaneous linear regressions, 286
 Singer, B., 457, 732, 1161
 Singer, E., 495, 742, 862
 Singh, M. P., 783
 Singleton, A., 761
 Sitjima, K., 532
Skewness, 543–544, 1137, 1159, 1193
 Skin cancer prevention, model of, 692
 Skinner, C., 157
 Skinner boxes, 640
 Sklodowska, Elzbieta, 1119
 Skoudis, E., 864
Slope, 231, 1190
 Smale, S., 730
 Smart, J. C., 1114, 1115
 Smirnov, N. V., 737
 Smith, A., 774
 Smith, A. F. M., 54, 55, 58
 Smith, E. R., 40
 Smith, L. M., 92
 Smith, Mary Lee, 636
 Smith, N. W., 887
 Smith, Peter W. F., 440
 Smith, S. K., 362
 Smith, T. M. F., 157
 Smith, T. W., 773
 Smithies, Christine, 1198
 Smithson, Michael, 412
 Smola, A. J., 727
 Sneath, P. H. A., 115, 116, 128
 Snedecor, W., 365
 Snijders, Tom, 390, 461, 653, 677, 724
 Snow, D., 561
Snowball sampling, 721, 771, 884

- Sobel, M. E., 609
 Social choice rule, 24–25
 Social constructionism, 183–185, 893, 1165
Social desirability bias, 641, 1183
 Social interview, 698
Social Justice and the City, 257
 Social network, 719
 Social relations model, 291
 Social Science Citation Index, 664
 Social Science Research Council, 565
 Social Sciences, National Research Council
 for the, 234
 Social structures, cognitive, 721
Social Theory and Social Structure, 643
Socially desirable, 1117
Society, 112
 Society for the Scientific Study
 of Subjectivity, 887
 Socioeconomic status (SES), 8
Sociological Imagination, The, 508
 Software, 553
 ATLAS.ti, 39, 87
 BMDP, 74–75, 668
 BUGS, 433
 CAQDAS (Computer-assisted qualitative data analysis), 87–89
 choosing, 88
 computer simulation, 158
 GLIM, 429
 latent class analysis, 732
 Mplus, 676
 N6, 705
 NVivo, 87–88, 748
 ordinary least square (OLS) estimator, 776
 partial least squares regression, 795
 Qualrus, 87
 relational database, 239
 SAS, 668, 676
 SPSS, 668, 676
 STATA, 632
 statistical, 402
 STATXACT, 1190
 Sokal, R. R., 115, 116, 128
Sorensen, Jesper B., 1177
 Spanos, A., 295
 Spearman, Charles, 39, 881
 Spearman-Brown correction formula, 512–513
 Spearman's rank-order correlation coefficient, 627
Specification, 170–171
 Spector, Paul E., 40, 41, 573, 641
 Speed, Terry P., 439
 Spence, M. T., 1180
 Spencer, Herbert, 641
 Spline functions, 422
 Spradley, J. P., 493
Spread, 1159
 SPSS, 28, 37, 236, 239
Spurious relationship, 1160
 Spuriousness, 302
 Srba, Frank, 313
 SSBG (sum of squares between groups), 321
 SSTOT (total sum of squares), 321
 Stable population model, 562
 Stainton Rogers, R., 887
 Stallard, E., 457, 1161
 Standard coordinates, 71
Standard deviation, 60, 64, 76, 260–261, 743, 895, 909, 1114, 1119, 1156, 1159, 1160, 1176, 1205
Standard error, 97, 535, 797, 897, 909, 1206
 Standard normal distribution, 744, 1167
 Standard query language (SQL), 239
Standard score, 1167, 1205–1206
Standardized, variable, 746
Standardized regression coefficient, 1167
Standards for Educational and Psychological Testing, 178, 179, 181, 213
 Standing, Lionel G., 451
 Standpoint epistemology, 381
 Stanley, J. C., 354, 502, 503, 515, 516, 618, 819, 899
 Stanley, Liz, 45
 Starting principle, 33
 Stata, 236
 STATA software, 632
 Stationarity. *See* **Cointegration; Time-series data; Trend analysis**
Statistic, 76, 319, 1155, 1159, 1176
 bias, 65
 decision theory, 270
 descriptive, 1159
 inferential, 435–436
 likelihood ratio, 440
 network analysis, 723–725
 nonparametric, 414
 official, 760–762
 test, 1148
Statistical inference, 38, 75–78, 776–777, 816, 869, 896, 1151
Statistical power, 614, 899
 Statistical regression, 503
Statistical Science, 112
Statistical significance, 827, 1125, 1202
 STATXACT software, 1190
 Staudte, R. G., 1143
 Stem-and-leaf plots, 543
 Stephensen, William, 887
 Stereotype, physical attractiveness, 451

- Stern, H. S., 433
 Stevens, Gillian, 250, 572
 Stewart, D. K., 85
 Stigler, S., 54, 61, 745
 Stillman, Todd, 645
Stochastic, 865
 form of **polynomial equation**, 832–833
 modeling, 904
 processes, 258
 time-series models, 22, 1128–1130, 1155, 1194
 trend, 1140
 Stokes, S. Lynne, 520
 Stoll, David, 1119
 Stone, C., 890
 Story grammars, 186–187
Strangers to Ourselves, 779
 Strassen's algorithm, 9
Stratified sample, 906, 1192
 Strauss, Anselm, 137, 180, 441, 442, 716, 724, 799, 1121, 1122–1123
 Strauss, David, 721
Street Corner Society, 386, 799
 Structural analysis, 706–707
Structural equation modeling (SEM), 11, 102, 164, 446, 802, 821, 901
 Structural linguistics, 382
Structuralism, 749, 846, 1124, 1183
 Structuralist research, 382
 Structuration theory, 642
Structure of Scientific Revolutions, The, 749, 839
Structured interview, 1143
Structured observation, 751, 752, 755
 Stuart, A., 622, 791
Studies in Ethnomethodology, 332
 Successive intervals, method of, 1126
 Suchman, Edward, 337
 Suci, C. J., 41
 Suczek, B., 1123
 Sudman, S., 773, 906
 Sudnow, D., 333
 Suen, E., 119
 Suen, H. K., 752
Suicide, 643
Sum of squared errors, 731
 See also **Residual sum of squares**
 Sum of squares between groups (SSBG), 321
 Summated rating scale, 41
 Summative evaluation, 339
 Summed deviance, 430
 Support vector machine (SVM), 727
 Survey data, 238
Survey Methodology, 112
 Survey Research Center, 896
Surveys, 853, 1162
 cross-sectional, 42
 data collection, 159–161
 design, total, 1135–1136
 Internet, 504–505, 766, 902
 Kish grid use in, 538
 panel, 42, 250
 pooled, 834
 research, 1115
 telephone, 740, 884, 905, 1116–1118, 1191
 Survival, duration, 342
 Survival analysis, 155
 Swanson, D. A., 362
 Swanson, J. M., 127
 Swedberg, Richard, 645
 Sweeney, Kevin J., 459
 Swicegood, C. Gray, 589
 Sybase, 239
Symbolic interactionism, 714, 749, 1123
Symmetric measures, 28
Symmetry, 1157, 1159
System of Logic, A, 415, 643
Systematic error, 1158
 Systematic introspection, 306
Systematic reviews, 831
Systems dynamics simulations, 159

T-distribution, 1114
T-statistic. *See* **T-distribution**
T-test, 76, 540, 1145–1147, 1192, 1194
 nonparametric statistics, 735–736
 t-distribution and, 1114
 Tables, random number, 911–912
 Tacq, Jacques, 271, 367, 669
 Tactical authenticity, 44
 Takane, Y., 670
Tales of the Field, 327, 1198
 Tamhane, A. C., 690
 TANF, 237
 Tannenbaum, P. H., 41
 Tashakkorri, A., 680
 Tassinari, L. G., 882
 Tau. *See* **Symmetric measures**
 Tau coefficient, Kendall's, 414
Taxonomy, 305, 535, 1113
 Taylor, Charles, 779
 Taylor, S., 185
 Tayman, J., 362
Tchebechev's inequality, 1113–1114
Team research, 1114–1115
Tearoom Trade, 799
 Teddlie, C., 680
Telephone surveys, 740, 774, 884, 905, 908, 1116–1118, 1191

- Telescoping, forward, 251
- Teller, A. H., 613
- Teller, E., 613
- Temporal understanding, 455
- Tenenhaus, M., 792, 794
- Tesluk, P. E., 1115
- Test-retest reliability**, 164, 525, 1119–1120
- Test statistic, 1148
- Testimonio**, 1118–1119
- Testing, 503
- Tetrachoric correlation. *See* **Symmetric measures**
- Text**, 846, 1120–1121
- qualitative content analysis** of, 889–890
- Thematic analysis, 186, 706
- Thematic Apperception Test (TAT), 874
- Theoretical autocorrelation (ACF), 667 (figure)
- Theoretical questions, 462–463
- Theoretical sampling**, 444, 606, 884, 1121–1122
- Theoretical saturation**, 1122–1123
- Theoretical Sensitivity*, 441
- Theoretical triangulation, 1142
- Theory**, 1, 17, 33, 34, 486, 1123, 1171, 1181
- catastrophe**, 94–96, 730
- chaos**, 120–121, 730, 850
- critical**, 223–224, 376–377, 1123
- critical race**, 220–221
- game, 27
- grounded**, 17, 21, 39, 88, 91, 127, 180, 440–444, 606, 706, 747, 748, 771, 799, 812, 892, 1121, 1122, 1123
- item response**, 881
- nature of, 92
- queueing**, 904–905
- testing, 149–150
- Theory-driven evaluation, 340
- Thick description**, 328, 1124
- Third-order. *See* **Order**
- Thom, R., 94
- Thomas, H., 389
- Thomas, J., 216
- Thomas, W. I., 467
- Thompson, B., 145, 147
- Thompson, J., 217
- Thompson, Paul, 770, 771
- Thompson, W., 861
- Thomson, A., 770, 771
- Thorne, S. E., 892
- Three-way ANOVA**, 1150
- Threshold effect**, 1124–1125
- Thrice Told Tale*, A, 1198
- Thurstone, Louis L., 40, 370, 372, 1125
- Thurstone scaling**, 40–41, 1125–1127
- Tiao, G. C., 22, 55, 80
- Tibshirani, R. J., 423, 857
- Time-contingent responding, 346
- Time diary. *See* **Time measurement**
- Time dimension, 476
- Time measurement**, 1127–1128
- Time-series cross-section (TSCS)**
- models**, 1128–1132
- Time-series data (analysis/design)**, 22, 1131–1134, 1172, 1173, 1190, 1194
- autoregressive, 22, 23
- estimation, 1129–1130
- examples of, 1129
- integrated, 22, 23
- interrupted**, 898–899
- moving average, 22, 23
- nonstationary, 1131
- panel, 1131
- specification and interpretation, 1130–1131
- unit root** and, 1158–1159
- Timm, N. H., 208
- Tinbergen, N., 336
- Ting, Kwok-fai, 491, 492
- Tinsley, H. E., 512
- Tobin, James, 109, 1134
- Tobit analysis**, 1134–1135
- Tolerance**, 1135
- Tomaka, J., 882
- Tommerup, Peter, 777
- Tootbaker, L. E., 690
- Total sum of squared variations (TSS), 625
- Total sum of squares (SSTOT), 321
- Total survey design**, 1135–1136
- Toulmin, S. E., 34
- Tourangeau, R., 823, 903
- Townsend, P., 761
- Transcendental idealism (Kant), 473
- Transcription, transcript**, 219, 1136–1137
- Transdiscipline, 338
- Transfer function. *See* **Box-Jenkins modeling**
- Transferability, 43, 1145
- Transformations**, 543, 779, 833, 1137–1138, 1205
- linear, 744
- Transition probabilities, 612
- Transition rate**, 1138–1139
- Transmission error, 1180
- Travels with Ernest*, 1198
- Traxel, Nicole M., 391
- Treatment**, 153, 898, 907–908, 910, 1139
- Tree diagram**, 1139–1141
- Trend analysis**, 1140–1142
- Treponema pallidum*, 105
- Triangulation**, 698–699, 755, 758, 1142–1143
- Trice, A. D., 495
- Trimming**, 403, 1143–1144
- Triplett, N., 899
- Trivedi, P. K., 342, 830

- Troubling the Angels*, 1198
- True score**, 1144
- Truman, Harry, 825
- Truncation, 277, 426
- See also* **Censoring and truncation**
- Trussell, J., 110, 1161
- Trust, 2–3
- Trustworthiness criteria**, 43, 894, 1144–1145
- Truth and Method*, 579
- Tsiatis, A., 156
- Tuchman, Gaye, 463
- Tucker, L. R., 70
- Tukey, John W., 52, 81, 273, 359, 360, 535, 687, 689, 690, 1137, 1143
- Tuman, N. B., 156, 453, 454
- Turrisi, R., 500, 696
- Tuskegee Syphilis Study, 18
- Tversky, A., 79, 1155
- Tweedie, Richard, 388
- Twenty statements test**, 1147–1148
- Two-sided test**, 10, 781, 1148–1149
- Two-stage least squares (TLS or 2SLS)**, 1149
- Two-state conditional maximum likelihood estimation (2SCML)**, 168
- Two-tailed testing. *See* **Two-sided test**
- Two-way ANOVA**, 13, 15–16, 764, 1149–1151
- Type I error**, 75, 781, 1151, 1152
- Type I sum of squares, 1153
- Type II error**, 1151–1152, 1202
- Type II sum of squares, 1153
- Type III sum of squares, 1153
- Types of models, 658
- Unbalanced designs**, 1153
- Unbiased**, 62–63, 788, 1154, 1179, 1191, 1200
- See also* **Bias**
- Uncertainty**, 1154–1156
- Understanding, 454–456
- Uniform association. *See* **Association model**
- Uniform Crime Reports, FBI, 447
- Unimodal distribution**, 1156–1157
- Unit of analysis**, 1157–1158
- Unit root**, 1158–1159
- United States Census, 108–109, 110–113, 760–761, 766
- Units of analysis, 673–674
- Unity of scientific method, 837
- Univariate analysis**, 73, 1137, 1159–1160
- Unobserved heterogeneity**, 1160–1162
- Unobtrusive measures**, 1162–1163, 1164
- Unobtrusive methods**, 1142, 1163–1166
- Unstandardized**, 1166–1168
- Unstructured interview**, 21, 537, 798, 1168–1169
- Urbina, S., 182
- Urry, J., 222
- Utilitarian theories, 323
- Utilization-focused evaluation**, 339, 1169
- V. *See* **Symmetric measures**
- Valentin, D., 727
- Validity**, 163–165, 216, 711, 716, 894, 899, 1171–1172
- consequential**, 178–179
- construct**, 182–183, 881, 882, 899, 1171–1172
- content, 1171–1172
- convergent**, 1143
- criterion-related**, 213–214, 1171
- data, 771
- external**, 25, 43, 712, 1144
- in **observational research**, 755
- internal**, 25, 43, 355, 618, 899, 1144
- key informant**, 537
- triangulation** for checking, 1143
- trustworthiness criteria** and, 1144
- Vallet, Louis-André, 655
- Value, median, 81
- Value judgments, 836
- Van de Pol, Frank, 612
- Van der Ark, L. A., 548
- Van der Heijden, P. G. M., 548
- Van der Wende, J., 154
- Van Dijk, Teun, 215
- Van Katwyk, P. T., 40, 41
- Van Leeuwen, Theo, 215
- Van Loan, G. F., 616
- Van Maanen, John, 327, 777, 841, 1197, 1198
- Van Manen, M., 580
- Vapnik, V. N., 727
- VAR (vector autoregression), 313
- Variable**, 445, 605, 770, 910, 1159, 1172
- continuous**, 193, 869, 898
- dependent**, 101, 150–151, 711, 1124, 1167, 1176, 1195, 1199, 1201
- determining codes for, 133–135
- Dichotomous, 7, 38, 1199
- dummy**, 11, 1137
- endogenous**, 102, 739, 849, 1149
- exogenous**, 102, 739, 849, 1149
- independent**, 101, 541, 1124, 1135, 1167, 1173, 1199, 1201
- mean of a, 1199
- nominal**, 11, 12, 37, 96, 728, 897, 1139, 1195
- observed frequencies and**, 757
- omitted**, 763
- ordinal, 37, 1199
- parameter models**, 1173
- polytomous**, 834

- predetermined**, 849–850
predictor, 725, 852–853, 1175, 1177, 1202
proxy, 878
qualitative, 894–895
quantitative, 11, 12, 64, 65, 894, 897–898, 1139, 1195
random, 39, 179, 715, 787, 1113, 1155
 trending, 1141
 Variance. *See* **Dispersion**
Variance component models, 762, 1173–1174
Variance inflation factors (VIFs), 1174–1176
Variances, 742, 773, 847, 902, 1113, 1114, 1151, 1167
Variation, 1176–1177
 (coefficient of), 1177–1178
 negative case, 716
 ratio, 124, 1178
Varieties of Religious Experience, 466
 Varimax rotation. *See* **Rotations**
 Vaupel, J. W., 457, 1161
 Vazquez, C., 874
Vector, 1178–1179
Vector autoregression (VAR), 313
Venn diagram, 169–170, 1179–1180
Verbal protocol analysis, 1180–1181
Verification, 216, 836–837, 1181–1182
 Vermunt, Jeroen K., 98, 100, 549, 553, 555, 581, 653, 666
Verstehen, 415, 454, 466, 1182–1183
 Victorian Database Online, 766
 Vienna circle, 377, 585–586
Vignette technique, 1183–1184
Virtual ethnography, 842, 1184–1185
Visual research, 1185–1186
Volunteer subjects, 26, 1186–1187
 Von Glaserfeld, Ernst, 185
 Von Thurn, D., 862
 Von Wright, Georg Henrik, 148
 Vuong test, 650
 Wachter, K. W., 110
 Wahba, G., 423
 Wahba, S., 615, 875
 Wald statistic, 588
 Wald test, 437, 676
Wald-Wolfowitz test, 1189–1190
 Walker, Robert, 19
 Walkerdine, V., 879
 Wall, D., 710
 Wallace, R. B., 864
 Wallnau, L. B., 1199
 Walsh, Susan
 Walster, E., 451
 Walton, J., 149
 Wan, C. K., 500
 Wand, Jonathan
 Wang, Cheng-Lung, 483, 609, 616, 815
 Ward, A., 772
 Warren, Carol A. B., 524
 Wasserman, L., 857
 Wasserman, Stanley, 720, 725, 1140
 Watkins, J. W. N., 642
 Watson, John B., 60
 Watson, M. W., 142
 Wealth indicators, 632
 Weaver, L., 800
 Webb, E., 1142, 1162–1166
 Webb, J. F., 831, 901
 Weber, Max, 152, 282, 415, 456, 466, 471, 473, 509, 728–729
 Wedderburn, R. W. N., 423, 426, 429
 Weibull model, 453
Weighted least squares, 459, 1190
Weighting, 76, 742, 1191–1192
 Weinberg, S., 36
 Weiner, C., 1123
 Weiner, L., 389
 Weisberg, Herbert F., 547, 1178
 Weisfeld, Carol Cronin, 336
 Weiss, D. J., 512
 Weiss, H. M., 40
Welch test, 1192–1193
 Weller, S. C., 139, 493
 Wengraf, Tom, 48, 69, 70
 West, S. G., 901
 Wetherell, Margaret, 185, 265, 507
 Wheldell, K., 751
 Whewell, William, 468
 White, D., 1180
 White, G. C., 87
 White, Harrison, 720
 White heteroskedastic consistent estimator, 251
White noise, 518, 666–667, 1194
 Whitehead, M., 761
 Whiteley, P. F., 80
 Whiting, John, 755
 Whitmore, E., 801
 Whitten, Guy
 “Who am I?” test, 1147–1148
 Whyte, William Foote, 386, 799
 Widrow-Hoff learning rule, 726, 727
 Wilcox, Rand R., 260, 261, 274, 275, 321, 690, 1143, 1144
 Wilcoxon-Mann-Whitney test, 627
 Wilcoxon rank-sum test, 606
Wilcoxon test, 541, 737, 1194–1195
 Wilks’s lambda, 271, 703

- Willer, David, 1
 Williams, E., 154
 Williams, Gareth, 706
 Williams, L. J., 641
 Williams, Malcolm, 421, 468, 474, 589, 642, 643, 712, 713, 769
 Wilson, Peter, 756
 Winch, P., 474
 Windleband, Wilhelm, 728–729
 Winer, B. J., 1192
WinMAX, 1195
 Winship, C., 251
 Winsorized variance, 275, 690
 Wishart distribution, 208
 Wisniewski, Wladek, 565
 Within-method triangulation, 1142
Within-samples sum of squares, 1195–1196
Within-subject design, 154, 196, 1196–1197
 WLS (weighted least squares), 459
 Wold, H., 792
 Wolf, Margery, 1198
 Wolfe, J. H., 554
 Wolfe, M., 861
 Women as subjects of research, 45, 771–772, 778, 1169
 Wood, B. Dan, 498, 506, 575
 Woods, James, 450, 486, 672, 696
 Wooldridge, J. M., 475, 1133, 1154
 Word counts, 186
 World wide web, 765
 Wright, S., 200, 802
 Wright, Sewall, 439
Writing, 1197–1198
Writing Culture, 661
 Wu, L. L., 454, 1139
 $\bar{X}$, 1199–1200
X variable, 763, 1199, 1201
 association and, 28
 baseline and, 52
 Xie, Yu, 32, 815

Y-intercept, 1201–1202
Y variable, 763, 1202
 association and, 28
 baseline and, 52
 Yahoo!, 765
 Yasuda's index, 653
 Yates, S. J., 185
Yates' correction, 1202
 Yeo, Stephen, 313
 Yin, R., 91
 Yin, R. K., 584
 Young, F. W., 670
 Young, K., 831
 Yuen, K. K., 1193
 Yule, G. Udny, 96, 759
Yule's Q , 191, 759, 1203–1204

Z-score, 61, 1205
Z-test, 1206–1207
 Zadeh, Lotfi, 412
 Zanna, M. P., 484
 Zavoina, W., 879
 Zeeman, E. C., 94
 Zellner, A., 425
 Zeng, L., 93
 Zero condition, 1154
 Zero-order. *See* **Order**
 Zimmerman, Marc, 1119
 Znaniecki, Florian, 17, 466–467
 Zola, Emile, 1197
 Zorn, C. J. W., 345, 424